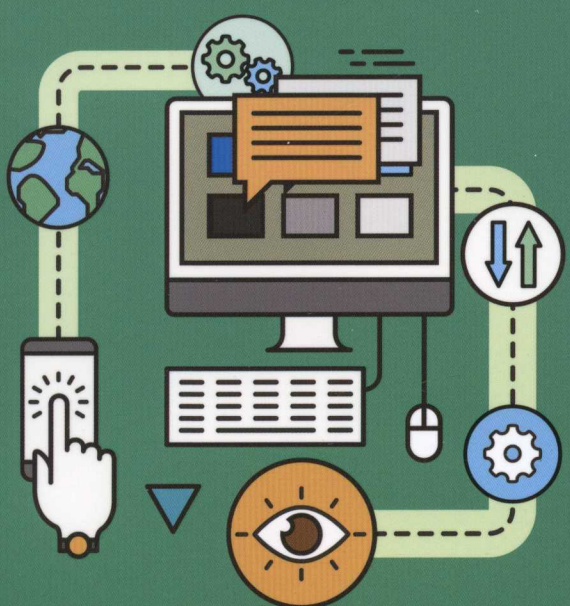
| Alistair Croll | 车品觉 | 宋星 | 曹政 | 吕厚昌 | 王淮 | 联合力荐

Broadview®
www.broadview.com.cn

数据驱动
从方法到实践

桑文锋 / 著



中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

作者简介

桑文锋

神策数据创始人兼CEO，浙江大学计算机科学与技术专业硕士，在百度任职8年，从无到有构建了百度用户日志大数据平台，覆盖数据收集、传输、元数据管理、作业流调度、海量数据查询引擎及数据可视化等。历任软件工程师、高级软件工程师、项目经理、高级项目经理、技术经理，2015年4月离职创建神策数据，针对企业客户推出用户行为分析产品——神策分析，帮助企业实现数据驱动。2017年7月，桑文锋荣获第六届中国财经峰会“2017最佳青年榜样”荣誉。

此外，神策数据联合创始人兼CTO曹犟，神策数据联合创始人兼首席架构师付力力，神策数据资深算法工程师邹雨晗，神策数据架构师房东雨，神策数据算法工程师韩越，神策数据数据分析总监陈新祥，神策数据用户行为洞察研究院负责人张乔，以及神策数据分析师高娜、薛创宇、李金霞、朱静芸均参与了此书的写作。

数据驱动

从方法到实践

桑文锋 / 著

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内容简介

本书是从理论到实践的全面且细致的企业数据驱动指南，从作者的百度大数据工作说起，完整还原其从零到一构建百度用户行为大数据处理平台经历。详解大数据本质、理念与现状，围绕数据驱动四环节——采集、建模、分析、指标，深入浅出地讲述企业如何将数据驱动方案落地，并指出数据驱动的价值在于“数据驱动决策”、“数据驱动产品智能”。最后通过互联网金融、电子商务、企业服务、零售四大行业实践，从需求梳理、事件指标设计、数据接入阶段、实际应用四大阶段介绍数据驱动在不同领域的商业价值，全面展示大数据在各领域内的应用情况与趋势展望。

本书贴近企业真实场景，兼具权威性与前瞻性，是广泛适用的普及读物，适合对大数据、数据驱动感兴趣的企业高管、决策者、创业者、IT人员、营销人员、产品经理、相关专业的学生等。

图书在版编目（CIP）数据

数据驱动：从方法到实践 / 桑文锋著. —北京：电子工业出版社，2018.3
ISBN 978-7-121-33451-1

Ⅰ.①数… Ⅱ.①桑… Ⅲ.①数据处理－研究 Ⅳ.①TP274

中国版本图书馆CIP数据核字（2018）第002302号

策划编辑：符隆美
责任编辑：张春雨
印　　刷：三河市华成印务有限公司
装　　订：三河市华成印务有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路173信箱　邮编：100036
开　　本：720×1000　1/16　印张：13.5　字数：260千字
版　　次：2018年3月第1版
印　　次：2018年4月第2次印刷
印　　数：3001～4500册
定　　价：49.00元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至zlts@phei.com.cn，盗版侵权举报请发邮件至dbqq@phei.com.cn。

本书咨询联系方式：（010）51260888-819，faq@phei.com.cn。

推荐语

文锋分享了他在商业数据的真知灼见，不盲目舶来，他明确地知道哪些理论在国内是行不通的，并传递出更本土化的理论。本书的结构和内容都经过了反复打磨，无论是从技术严谨性，还是从内容的实用性上看，都堪称互联网商业数据的可贵佳作。

——宋星，互联网数据官创始人、网站分析在中国创始人

数据的价值在哪里？作者根据其丰富的百度经历以及与众多客户深度碰撞后的思考，从方法论的高度全链路定义了数据采集、数据建模、数据分析与指标四大关键环节，并以实践诠释了如何用数据驱动决策、产品和业务，值得读者细细品味。

——赵军科，百联大数据总监

得益于文锋深厚的技术背景和丰富的实践经验，这本书清晰剖析了从采集、建模到分析运用的数据驱动全链条，值得每个数据人阅读。

——赵祺，今日头条增长团队负责人，前车来了联席 CEO

在不远的将来，不管你处在什么行业什么职位，数据分析都是你不得不具备的一种能力。本书提供给你一个极好的知识储备的机会，它有三点非常值得推荐：第一，浅显易懂地表达大数据的底层技术，让你能够明白数据怎么产生，怎么加工，怎么存储和运算；第二，抛开了晦涩难懂的各种模型和算法，将最普适的数据洞察和分析的方法呈现给你，让你能迅速具备“阅读数据”的能力；第三，清晰地将电商、互联网金融、零售、SaaS 软件等行业鲜活的数据应用案例呈现给你，让你加深对数据应用的理解。

——胡晨川，《数据化运营速成手册》一书作者，饿了么数据专家

文锋在百度的经历积累了大量本土化的业务实战经验，这本书浓缩了他近十来年宝贵经验的精华，一如神策分析的诞生，对于整个行业来说都是值得欣喜的事情。神策数据快速武装企业的数据部门，快速积累数据，并让所有在践行数据驱动业务增长的企业，都可以快速上路，让数据驱动最终成为每个公司的“标配”。

——刘晨，纷享销客联合创始人，数据中心总经理

随着大数据和智能时代的来临，数据驱动必然会变成人人都要具备的能力。本书里面的每条经验，都是一场场实战打出来的。与很多纸上谈兵的文字不同，本书的实例信手拈来，可想而知经历多少次的打磨才能有这样的效果。这使得本书内容对实际工作有着很强的指导作用，适合每个与数据打交道的人，常读常新。

——孙文亮，杏树林数据总监

作为数据驱动在初创公司的实践者，我们经历了从手动跑数据分析的“石器时代”到实时数据分析系统的“蒸汽时代”。工具已经成熟了，但在方法层面自己则一直瞎练野拳。一见到本书，就有相见恨晚之感，数据驱动终于有了成体系的“招式”！屠龙宝刀，要配上好武学，希望本书可以帮助更多公司实现数据驱动。

——黄震昕，造数科技创始人兼 CEO

推荐序 1

If companies were people, then we would be in the middle of one of the greatest health crises of the modern age. Once, the lifespan of a company on the Fortune 500 index of large businesses was 65 years. Today, it's only 20. In the last decade, most of the world's big, reliable firms have been displaced by digital upstarts: Apple, Amazon, Tencent, Google, Baidu, and Alibaba.

It gets worse. The chances of a company reinventing itself are low. The Corporate Strategy Board says efforts at digital transformation fail 95% of the time; Clayton Christensen, author of *The Innovator's Dilemma*, puts the number at 99%.

But there's some good news, too. Because technology has given us the ability to measure everything, accurately, better than ever before. We can know ourselves.

A 2011 MIT study[1] found that companies that use data-driven analytics instead of intuition have 5%-6% higher productivity and profits than competitors. Over a few years, data and analytics is the difference between success and obscurity.

Data, it is often said, is the new oil. Data replaces opinions with accuracy, letting us know our customers, our suppliers, and ourselves with unprecedented clarity. And data is the food of artificial intelligence, because it's how we train machine learning algorithms.

---

[1] Brynjolfsson, Erik,LorinHitt, and Heekyung Kim. "Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?" Available at SSRN 1819486 (2011).

On its own, oil isn't very useful. It just sits in the ground. To put oil to work takes an ecosystem: Refineries, gas stations, motors, regulations, roads, and more. And so it is with data. Simply collecting it won't help you; you need to extract it, clean it, analyze it, execute on what you learn, and feed that learning back into your systems.

As technology replaces many traditional tasks through automation and machine learning, we may wonder what is left for humans to do. The answer is simple: Think critically about what we want those machines to do for us. The most important skill for a human, whether they're a startup, an analyst, or a manager, is to ask the right question.

Asking good questions is harder than it seems. It requires an understanding of the existing business model, the competitive landscape, and the resources at your disposal. But it also requires that we know that the existing business model is outdated, vulnerable, and ready for change.

A world powered by real-time information creates two roads. One road is littered with the bodies of companies that couldn't make the transformation, preferring anecdote over fact. The other road is paved with the profits of those who learned to harness data and embrace analytical thinking.

You're at a fork in this road. And right now, you're holding the map that will steer you down the right path.

如果我们将企业比作人类，那么许多企业正处于壮年的巨大健康危机之中。曾经，世界 500 强企业的生命周期是 65 年，而现在仅有 20 年。近 10 年来，众多规模大、可靠的企业已被“数据新贵企业”所替代，例如苹果公司、亚马逊、腾讯、谷歌、百度、阿里巴巴等。

更糟糕的是，企业进行自我重塑的概率变得越来越低。公司战略委员会指出，95% 的企业数字化的转变是失败的。《创新者的窘境》一书的作者克雷顿·克里斯滕森认为这一数字已达到 99%。

当然也有好消息：科技赋予我们衡量一切事物的能力，我们能更好地认识自己。在这点上，曾经的任何时代都难以企及。

麻省理工学院的一项研究表明[1]，相比依靠直觉来实现决策的企业，那些通过数据驱动实现决策的企业拥有更高的生产效率和利润。这类企业的生产效率和利润普遍高于竞争对手 5% ～ 6%。显然，未来是否拥有数据分析能力，将决定一家企业是成功，还是逐渐销声匿迹。

我们经常说，数据是新石油。数据的准确性代替了“意见”的主观性，让我们更好地了解我们的供应商、我们的顾客以及我们自身。同时数据也是人工智能的基础，因为我们正是通过数据的运用来实现机器学习的。

对石油来说，一直被埋藏在地下的石油并无价值。它的价值在于应用，石油开采需要一个“生态系统”：炼油厂、加油站、汽车、规则、道路等。数据也是如此，仅仅收集数据并无价值，你需要提取、清洗、分析，让分析结果得以执行与运用，并反馈至“生态系统”中。

随着自动化操作和机器学习代替了部分传统工作，我们为此很疑惑：还有哪些工作需要人类来做？答案其实很简单：我们需要辩证地思考究竟人类需要机器来做什么。无论是初入职场的新人、分析师，还是企业管理者，提出正确的问题是他们最重要的能力。

但是，这实现起来很难。提问者既需要了解企业当前的商业模式、竞争格局以及可控资源，也需要意识到现有商业模式已经变得过时、不稳定，而且亟待改变。

信息随时随刻在产生，它为世界指出两条路：一条路布满着那些故步自封、因循守旧企业的“尸体”；另一条则为拥有数据思维和掌握数据驾驭能力的企业铺就康庄大道。而此时此刻，你正处于交叉路口，手中恰好握着一张指引正确路径的“地图”。

Alistair Croll

哈佛商学院访问执行官，Coradiant 公司联合创始人

《精益数据分析》一书作者

[1] 《数据驱动的决策是如何影响企业绩效的》社会科学研究网 1819486（2011 年）。

推荐序 2

数据驱动的概念已经被各个行业广泛认同，但认同与落实之间，还是有相当的距离，这里最大的障碍是，技术人员缺乏对业务的理解，而业务人员又无法理解和充分利用技术，有数据却用不好、不会用是很常见的弊病。即便是一些有数据分析、研发实力的企业，也面临从需求到实现的巨大研发成本和时间周期等问题，导致决策效率低，对瞬息万变的市场情况，无法做出快速有效的应对。

百度早期的技术资源有限，主要技术资源优先考虑产品研发迭代，对数据分析的支持力度不足。2005 年我参与创建百度的商业分析部门，因为无法得到充足的技术资源，只好自己动手，在产品部门架构内处理数据，解决业务诉求所需的数据分析，所幸那时候百度的业务数据规模有限，每日的部分业务数据日志尚处于 GB 级别，按照我们有限的技术能力，单服务器勉强可以应付。

2007 年之后，百度的业务规模急速扩大，业务部门也越来越重视数据决策方向的诉求，幸而此时技术资源也得到了有效的扩充，在桑文锋同学的有力支持下，百度的数据分析能力和整体架构都得到了翻天覆地的革新和发展，针对诸多核心产品升级，数据决策的意义和价值也得到了充分的彰显。

能解决一个巨头公司数据分析领域的技术瓶颈，提升数据决策能力，已经是一项了不起的成就，但文锋的目标显然不止于此，搭建一套通用灵活的技术架构，显然有更广阔的应用场景。让一线业务人员在不需要充分理解技术的前提下，快速针对业务诉求完成数据分析，实现数据决策，这是神策数据（Sensors Data）项

目的一个愿景。

我从百度出来后进入了游戏行业，后来辗转到海外发展，对国内行业的现状了解不多。说来也有意思，好几个游戏行业同行创业者，在不同场合主动跟我提及神策数据非常有价值，对他们的业务帮助很大，我才注意到文锋的创业项目，并钦佩于他们现在所取得的成就，这个成就，不是说这家公司收了多少服务费，赚了多少钱，而是他们真的有效提升了整个行业的数据决策能力，有效降低了数据决策的操作成本和门槛，这个价值是从业者们尤为要感谢的。

感谢文锋，提前让我阅读了这本书籍，我觉得，对于希望提升数据决策能力、了解数据决策真相的从业者，这本书是很好的读物，其内容并非晦涩难懂的技术描述，而更多是对数据驱动和数据分析的理解，并以亲身案例作为辅助讲解。建立正确的认识是做好数据决策的前提，而其中所提到的很多问题场景，相信也是很多从业者经常遇到和面对的。

以上，希望对您的阅读和选择，有所帮助。

曹　政

曾任百度商业分析部经理，现知名 IT 自媒体博主

互联网游戏出海领域创业者

推荐序 3

我一直觉得数据分析是一种修行，“修”的是思考的能力，“行”的是落实成为方案的方法。经过多年的工作，正是不经一番寒彻骨，怎得梅花扑鼻香。回想我与文锋初次见面便谈到数据化的过程，阿里与百度都经历过这样的挑战，我想这便是他请我写推荐序的原因吧。

以前企业中只有一小部分人具备数据分析的能力，随着近几年数据平台的成熟，数据从收集到使用越来越方便，以往想要出一份分析报告可能要等上数周的日子已经一去不复返。曾经有一位业务方代表对我说过，在等候分析报告出来与拍脑之间，我选择了后者，因为时机更重要。可想而知决策的速度很关键。在后信息时代，DT[1] 的普适度将变得更直接、简单。未来的智能时代，我很相信很多分析报告也将被自动化的智能决策所取代，届时智慧的人类也将要“升级”到“神策”的阶段，人更要学会驾驭决策上的决策、逻辑上的逻辑。

当然，理想归理想，在智能决策的路上还需要很多同行们努力，而文锋在书内的描述正是他这几年创业的发现与精华。

车品觉

红杉中国专家合伙人、全国信标委大数据标准工作组副组长

[1] DT，Data Technology，数据处理技术。

推荐序 4

当今物联时代，业界同仁都在谈大数据和人工智能。大数据已成众多公司的核心资产，大数据战略已成众多公司的核心战略。之所以如此，一是因为大数据技术的普及，二是因为大数据已经为无数企业带来了实打实的核心价值。大数据 4V[1] 中最重要的还是接地气的价值驱动——Value。使用大数据技术，挖掘大数据价值，不断优化用户体验、客户体验、产品体验，已然成为当今企业成功的金科玉律。

1996 年我在美国正式进入职场。我在职业生涯的早期就对数据情有独钟，那个时候还没有大数据这个提法。这不仅仅是出于对数据技术的喜好，也是因为我在做数据项目的时候，真正体会到了数据给业务带来的不同。1998 年我加入 Yahoo!，成为第一个专门做数据的工程师，用一句话总结我在 Yahoo! 7 年的工作，那就是使用大数据更好地理解用户，驱动用户产品创新，更好地服务用户。2005 年我离开 Yahoo! 加入 Google 是源于好奇心，当时 Google 的流量是 Yahoo! 的 1/10，但收益却跟 Yahoo! 一样多。为什么搜索会这么赚钱？用一句话总结我在 Google 6 年的工作，那就是使用大数据能更好地理解客户广告诉求，驱动广告产品创新，更好地服务广告主。

2011 年我有幸加入百度带领数据团队。百度是一个对大数据工作非常重视的公司。大数据工作是百度的核心竞争力之一，其核心搜索业务也是建立在大数据

[1] 4V，Volume（大量）、Velocity（高速）、Variety（多样）和 Value（价值）。

技术之上的。文锋是我在百度工作期间的爱将。在百度工作的几年中，我跟文锋、曹犟、力力、耀洲等聪明能干、充满活力的同学们一起，在实战中不断总结与学习，一同推进大数据技术的进步，这是一段非常享受并有成就感的经历。

我在百度大数据工作时，跟小伙伴们一起启动了不少项目，一切都围绕发挥大数据价值而发力。大数据价值从让数据说话开始，大数据驱动决策。几乎每一个产品都是一个闭环的生态。从产品上线的第一天起，用户就在不断用手或脚投票，告诉你哪里好用、哪里需要改进。用户越多，这个闭环正负反馈的信息量就越大。当我们可以快速地把这些信息以报表分析的形式，展现给我们的产品经理、产品研发工程师及各级决策者们的时候，就能不断地发现机会、迭代改进产品。当数据量达到一定规模后，数据所反馈的趋势就越清楚，这不仅体现在更好地理解现有需求上，也会不断挖掘新的需求，预测引导用户需求，不断改进创新产品。

搜索如此，广告如此，新领域创新也是如此。从预防疾病，提升百姓健康体验，到挖掘旅游热点，提供最佳出游体验，到因材施教，颠覆特权教育，到预测交通流量，改善交通拥堵，大数据驱动颠覆式创新。

大数据的另一个更重要的价值在于让数据为用户工作，驱动个性化服务。当数据量达到一定规模后，因人工智能算法已经普及，故对用户每一次产品使用背后意图的把握就会越来越精准，从而可以做到为用户提供有针对性的个性化服务。这种个性化可以从用户群组个性化开始，也就是对不同类型的受众提供不同的服务，可以做到针对每个用户的个性化服务，甚至细化到对每一个用户每一个动作的个性化服务。大数据价值在这一点上的发力可以真正引爆产品生态闭环的马太效应。

文锋在书中把他过去丰富的实践经验做了非常好的总结，干货满满，源于实践又高于实践。文锋一直想成为中国大数据产业兴旺的推动者，他创建了神策数据（Sensors Data），不断践行自己的理想。本书字里行间生动活泼，也体现出作者对大数据领域的理想情怀和脚踏实地的实干家精神。对大数据行业的每一位实践者和企业家来说，本书都非常值得一读。

吕厚昌（Alex Lu）

曾任百度高级总监，Pinterest 大数据部负责人

推荐序 5

我认识桑文锋是因为投资的事情。2015 年初的某一天，朋友给我介绍了一个人，说在百度做了很多年大数据基础架构，有丰富的实战经验，又是我浙江大学的学弟。这样稀少的人才，当然要见见。

我原来在 Facebook 做了很多年工程师，对数据驱动非常熟悉，也非常坚定地相信其价值。基于数据的决策就像船员在茫茫大海之中看到了灯塔，就像飞机飞行在迷雾之中但装有雷达。有时候凭经验拍脑袋也许有用，但有了地图的驾驶员，一定比最好的老司机更不容易掉坑里。Facebook 在这方面做了很多工作，用数据来辅助所有（没错，是所有）的产品决策。日志系统、ETL、Hadoop/Hive、实时的数据仪表盘、A/B 测试、灰度发布，这些琳琅满目的数据工具组成了一个套装，为 Facebook 在商业战场的迷雾之中提供了看清正确方向的“千里眼”和“顺风耳”。Facebook 最早做 Hadoop/Hive 的人就是我从 Yahoo! 推荐过来的。我在 Facebook 做过的产品包括 NewsFeed、Giftshop、SocialAds，无一不是深度应用数据的典型产品。我在 Facebook 的最后两年负责支付相关的数据平台和安全系统，这些工作更是对数据从头到尾都有很强的要求。Facebook 一向的实践是相信数据，但又不迷信数据。利用数据，但不只依靠数据。

但我在 2012 年回到中国的时候，发现数据驱动的理念和做法在中国没有太多的公司在实际操作。当时大多数公司，都还聚焦在粗放型增长，做产品主要靠拍脑袋，没有太多应用数据的工具和能力，更可惜的是，没有应用数据来指导决

策的意愿。少有的既懂理论又有实践的人，基本上在 BAT 这三家公司，尤其是数据技术利用最早的百度。

认识百度出来的桑文锋，在数据驱动这件事情上总算找到了知音。文锋的这本书，尝试去解决两个很有意义的问题。一是如何在思想上将原来拍脑袋决策的方式改变为用数据来辅助决策；二是如何让更多的公司更容易地获得数据驱动的能力。虽然我给很多公司做过分享，但我知道数据辅助决策的思想不会很快在中国的互联网公司实现，更何况有很多有数据而不知道怎么去用的传统企业。但桑文锋对于整个数据流程非常熟悉，例如，如何通过埋点获得数据，如何对数据进行结构化，如何对结构化的数据进行最优的存储和查询，如何将数据链条串起来进行最深度的分析，如何对数据做最好的展示以便更好地决策。在这一方面，他是我在中国见过的最有能力、信念最坚定的一个人。

我们相信桑文锋驾驭数据驱动商业的能力，也相信他身上那股坚定的信念，他愿意花很多年，付出很多努力，将数据基础能力像水和电一样提供给中国企业。我们将自己的资本和信心赌到桑文锋身上。我们也相信这本书，会给希望在商业战场上多一双数据眼睛的企业家很多帮助。

王　淮

《打造 Facebook》一书作者，线性资本创始合伙人

目　录

第 1 章

从百度大数据工作的经历说开

2007 年我从浙江大学研究生毕业，作为一名软件工程师正式加入百度搜索新产品部的百度知道研发团队。入职第一天，在成功登录邮箱账号之后，我发现邮箱里已经有数十封统计报表邮件，包含了产品的详细统计数据，如检索量、提问量、回答量等。从此，我正式开始了与数据打交道的历程。

百度文化中有一条是用数据说话。不管是产品经理的产品功能设计，还是功能上线后的效果评估，或者是工程师开发的模块性能，都需要用数据说话。如果没有数据支撑，方案就无法通过，因此百度公司有大量的统计分析工作。

我先后做了百度知道的待解决问题推荐、全百度日志统计平台、用户数据仓库、数据源结构化等一系列数据相关的项目。我带的团队，从最初三四个人的小团队逐步成长为百度公司的大数据团队，最后成为独立的大数据部门中的核心部分。在带领新产品部门的数据团队之初，我就给自己设定了目标：在百度公司内，我要让大数据团队发展成为一支与自然语言处理团队同样地位的团队。在我 2015 年离职前，基本达到目标。

数据工作不同于学习一门编程语言，也不同于在实验室里做理论研究，它需要大量的实践。许多人认为，大数据处理就是把一些开源组件拼凑到一起。但真正做起来后，就会发现人才储备、数据源、数据流建设、数据分析方法等一系列问题源源不断地冒出来。比较幸运的是，我算是在国内最早一批接触大数据的从业人员，并且积累了大量的实践经验。

百度数据板块：网页数据和用户行为数据

百度内部有两块重要数据：网页数据和用户行为数据。就网页数据而言，百

度在 2000 年做搜索业务时，全国中文网页数量不超过 2 亿个，从网页上整体抓取的数据只有几百 GB。谷歌从 1998 年开始做搜索，当时抓取了 2500 多万个网页的内容，压缩后只有 47GB。谷歌与百度这十几年来都在不断迭代，但经常被用户访问的部分已趋于稳定。最近几年，百度的常用网页库数据有几百 PB。

用户行为数据是指用户每次访问百度的产品所留下的痕迹。比如你在百度搜索上进行一次检索，就会在服务器上留下一条记录，记录了你的检索词、Cookie 信息、访问时间、IP 地址等。用户行为数据的条数比网页数据的高一个数量级，因为对于同一个页面来说，用户在页面创建时就会产生一次用户行为数据，而多次访问也会产生行为数据。在 2008 年我最开始做日志统计平台时，整个新产品部每天产生的行为数据有几十 TB。到我离开百度时，全公司每天采集到的用户行为数据达到 PB 级别，而现在只会更多。也就是说，现在几个月产生的用户行为数据，就可以和常用网页库数据相当，并且这些行为数据很有价值。

搜索引擎发展

搜索引擎的发展共经历了三大阶段，分别是内容相关性、网页链接关系和基于用户行为。

最初，所有的搜索引擎都基于关键词匹配相关内容，只要检索的关键词与实际网页内容匹配就可以显示。但是随着内容增多，排序就成了问题。同时，作弊现象也开始出现，如果将垃圾页面塞进去，网民根本无法搜索到有用的内容。

搜索引擎第二个阶段，即基于链接关系决定排序。当时，谷歌的拉里·佩奇、道琼斯公司的李彦宏，还有乔恩·克莱伯格教授，他们三人都意识到，链接本身很重要，一个网页被链接多少次，决定这个网页本身的权重是多少。

这也是谷歌、百度起家的技术，许多人以为百度抄袭了谷歌，但我通过研究李彦宏和拉里·佩奇的专利，发现李彦宏的专利是在 1997 年提交的，拉里·佩奇的专利是在 1998 年提交的，可见李彦宏还更早一些。

2005 年左右，搜索引擎进入第三个阶段 —— 基于用户行为。通过网页链接关系可以将高质量的网页排在前面，然而，随着新内容的不断增加，这种方式又暴露了新的问题。比如某个明星曝出了新绯闻，而一些老的内容具有更高的权重，被排在前面。在用户搜索时，会点击排在靠后位置的新内容。这样，搜索引擎就可以利用记录的“用户点击”数据，来实时调整结果页的排序，将点击更多的结

果排在前面。也就是说，通过用户行为数据，搜索引擎具有更好的效果。到目前为止，用户行为已经占据搜索引擎非常大的比重，根据一些业内专家的说法，用户行为权重已经超过50%。

用户行为分析践行：百度知道的回答量提升7.5%

在我刚加入百度时，“百度知道”已经成立三年，采用“问答”的形式，每天有9万多次提问和25万多次回答。由于产品形态成熟、数据稳定，所以优化与提升空间非常狭小。为了提升百度知道的回答量，我们开始研究用户，并尝试对不同用户采用不同的策略。比如为他们展示不同的样式和界面，以此来提升百度知道的产品黏性和价值。在2008年初，我们开始尝试通过待解决问题推荐的方式来提升回答量。

第一次，基于核心用户。我们抽取了35万个核心用户群——近1个月回答问题的次数在6次之上的用户群体，我们为该用户群体抽取了17万多个兴趣词并做了个性化推荐。前后历时3个多月，结果却十分令人失望。我们发现，用户将回答问题的入口从之前分类页面改为个人中心，仅此而已，用户回答量没有发生变化。

对此，我们进行了反思。一般来说产品的优化与提升只有两种思路，要么吸引更多新用户，要么在单个用户上“榨取”更多价值。既然老用户被“榨取”得差不多了，不妨尝试拉新用户，进而扩大用户规模。因此，我们进行了第二次尝试，基于所有用户做个性化推荐，而非仅针对核心用户。

百度内部当时有一个项目叫“后羿”，起源于百度在2008年做个性化广告的设想，即在用户进行搜索操作时，基于用户所搜索的关键词和用户行为记录，为用户推出相关广告。用户通过浏览器进行访问的时候，都会种下一个Cookie，用户在百度贴吧、百度知道、百度网页所浏览的信息都能通过Cookie串到一起。这为后续进行用户行为分析打下了坚实的根基。

于是，我们直接基于这些数据，根据用户的检索和访问页面的标题进行兴趣模型训练，然后抽取每个用户权重最高的5个兴趣词，当用户访问百度知道的详情页时，我们基于每个用户的兴趣词做实时搜索，将7、8个待解决的问题放到页面右侧。这次尝试效果非常好，新版上线后，百度知道的回答量提升了7.5%，而我也因此获得当时百度个人的最高荣誉——“最佳百度人”奖项。接下来，

我对百度知道又做了一些改良，比如让推荐问题更具多样性、按照用户对“兴趣点”发生的时间进行权重调整等。但我也发现再往后提升就比较困难了，在这之后，我被安排到一个数据统计团队工作。

从零到一构建百度大数据分析平台

从 2008 年加入数据统计团队之后，我就开始专注在大数据分析平台。当时还没有“大数据”的概念（大数据的概念大约在 2011 年出现），我在百度从零到一做这个事情的过程可以分成三个阶段。

第一阶段：2008 年，日志统计平台

2008 年，百度流量已经很大，尤其是百度知道、百度贴吧的数据量。前面提到，百度强调要用数据说话，这点我是非常认可的。百度做产品、功能都要基于数据。当我们需要进行流量统计和数据分析时，就遇到了问题。

因为各业务都会有处理起来非常烦琐的需求：要写脚本[1]、上线，需求响应时间很慢，整个需求周期非常长，维护多个脚本十分麻烦，很容易出问题。当时主要基于单机来计算，数据规模稍大的任务，通常要跑好几个小时。

为解决这个问题，我们当时想到使用 Hadoop。

可以说 Hadoop 是整个大数据生态的根基，其作用就像 PC 领域的 Windows。通过它我们可以实现海量数据的存储和分布式计算。当然，我们现在所说的 Hadoop 生态，还包括了数据传输、机器学习等其他组件。

当时 Hadoop 还只是测试版，使用起来非常不稳定。我们在进行平台设计时，留有两套计算接口：一套将数据提交到 Hadoop 平台，一套将数据提交到已有的单机服务。

Hadoop 到底能不能解决我们的日志统计问题，我们心里没底。如果 Hadoop 满足不了需求，我们就还是用单机做计算。做一个平台并不难，关键是怎么做一个好用的平台。我把常用的统计分析需求进一步抽象，分别抽象为计数统计、去重统计和 Top *N* 统计，并设计了一个界面，可以通过点选直接生成对应的任务，整个操作非常流畅。图 1-1 是当时我们做的日志统计平台架构图。

---

[1] 用 Perl、Python 等解释性语言开发的简单程序。

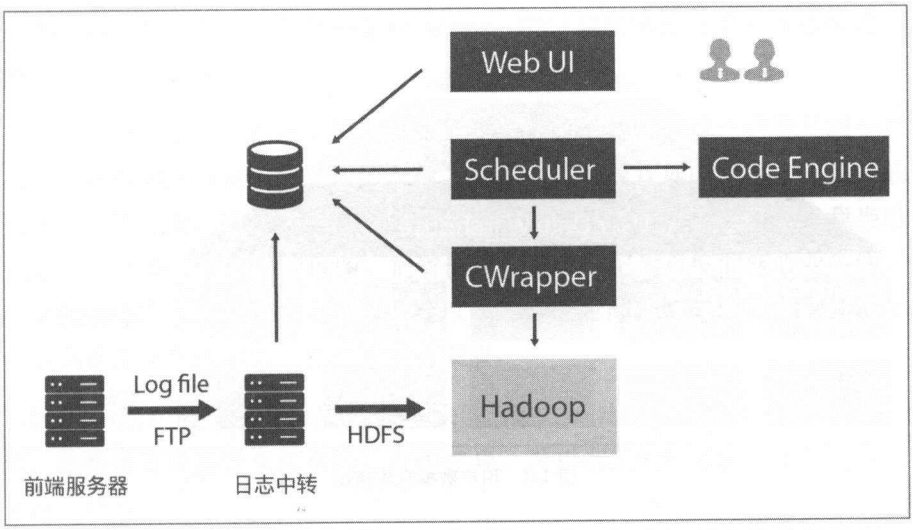


图 1-1　日志统计平台 LSP 1.0 架构图

平台发布后的效果让我很震惊。首先是常规的需求开发，从几天降到了几分钟。其次是运行周期，从单机计算变成一百多台机器分布式计算，几个小时的任务变成一两分钟。经过一年多的时间，整个公司都统一到这个平台。这是我在百度做的最有成就感的一件事。

但是，基本统计需求得到解决后，很多新需求又被释放出来。由于整个公司都在用，用于日志统计平台的机器从 100 多台增长到 5000 台，我们每个季度提预算的时候都要提 1000 台机器，我心惊胆战，毕竟日志统计团队做的这些统计任务到底有多大价值，很难衡量。

后来我的团队从以计算为中心的思路，转变为以数据为中心，也就是构建数据仓库。

第二阶段：2011 年，用户数据仓库

当时百度已经有几十条业务线，这些业务线从源头产生的数据质量不高，而且推动这些业务线进行改造实在太难了。我们就采用折中的方式：保持源头不动，将非结构化的数据结构化，使整个公司的业务线形成用户数据仓库，在这个基础上构建不同业务的主题数据，在此之上建立 BI 支持，这就形成了一个数据金字塔，如图 1-2 所示。

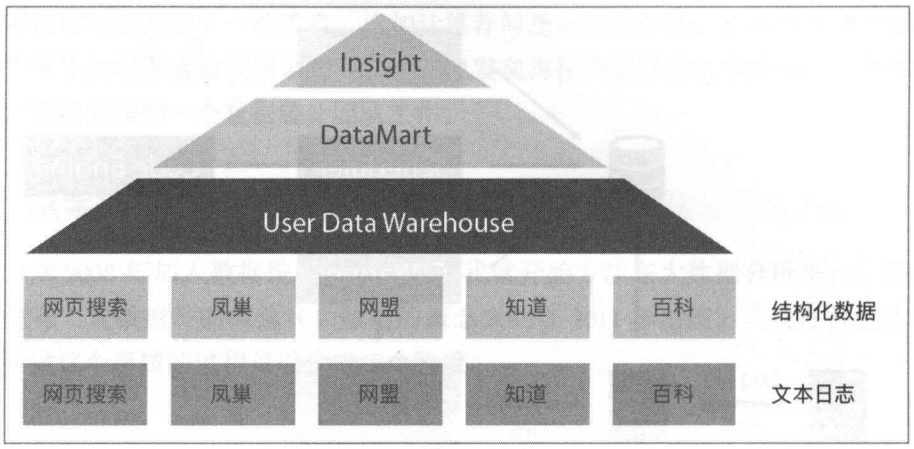


图 1-2　用户数据仓库模型

这其中最核心的就是 User Data Warehouse 部分。我们当时采用一种 Event（行为事件）模型，把用户在百度进行的任何一次行为记录，都规范为一个 Event。Event 的属性包括用户 ID、时间、设备信息、行为特有的参数等。这样，全百度的业务线都统一到一张表上，如图 1-3 所示。我们通过用户 ID 把用户在百度各个业务线的访问行为全部抽出来，在这上面做数据挖掘、数据分析变得非常容易。

用户ID	事件类型	时间	国家	省份	URL
ID01	注册	…	…	…	…
ID02	登录	…	…	…	…
ID03	搜索	…	…	…	…

图 1-3　用户行为事件

第三阶段：2013 年，数据源管理

当我们构建好整个数据金字塔，进入新的数据阶段后，又出现新的问题。虽然整个架子搭起来了，但是四处漏风。每次源头的变更，我们都要进行新的数据清洗和入库工作，开发周期和后续的运算周期非常长。业务线在上线之后不能马上使用数据，我们数据团队也疲于奔命。痛定思痛，我们觉得问题的关键还是在数据源，要从源头解决这个问题，如图 1-4 所示，数据源管理可以分成以下三块。

第一块是从数据源方面，将我们开发的内部结构化日志打印库和字段变更为审核系统，引入 Google Protocol Buffer 作为结构化的格式。

第二块是开发新的实时传输系统 Minos，将批量数据传输的方式改造为实时数据传输。

第三块是查询，对查询引擎本身做了改造，改造的时候提出数据从源头产生之后马上就能通过查询引擎分析的目标。在整个数据源管理的项目中，最难的不是系统组件的开发，而是推动各个业务线配合升级新的日志打印方式。我当时让成员做了一个 Web 版的中国地图，把省份和大城市标记为百度的核心业务线，每推动一个地方完成改造就插上红旗。经过一年半的时间，这份地图上都插满了红旗，这是我在百度做的第二有成就感的事情。

回想 2012 年，当时我和从 Google 来的总监上司沟通，他说 Google 源头产生数据，很快就可以进行 SQL 分析，我很诧异。没想到我们经过两三年的时间也达到了这一状态。

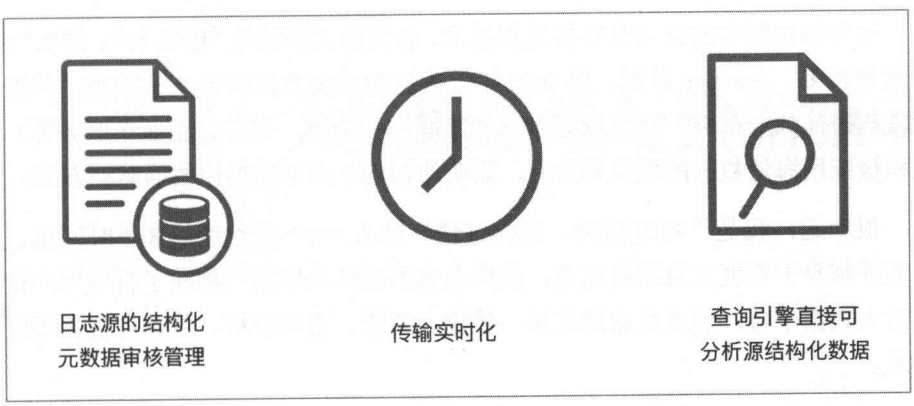


图 1-4　数据源管理

数据源与 Event 模型的重要性

总的来说，我在百度做用户行为数据平台的心得有以下两点。

1. 数据源很重要。若想把数据平台做好，数据源非常重要。例如，网页搜索是百度最核心的业务线，其他的都是附属业务线，这些业务线都会用到网页搜索的日志数据。如果变更搜索的日志格式，下游依赖搜索的业务程序可能都会瘫痪，但如果我们从源头本身结构化，下游就不需要跟着源头动，数据解析效率也会高很多。

这也是我们创业思路中的核心之一。在数据采集这一块要“大”“全”“细”

“时”。后续会对此观点详细介绍，此处不再赘述。

2. 用户行为事件模型很有效。规范并结构化用户行为之后，许多数据分析都会变得更容易。每个 Event 都是用户发生行为的一个快照，能够尽可能地还原现场。相比之前只是简单统计条数的访问量模型，Event 模型更精细化，这种模型的威力会远超想象，后续章节会详细介绍。

大数据是屠龙术

中国互联网化 20 年，经历了从“拍脑袋”到“数据驱动”的演进。从 2015 年至今，我国企业互联网化进入全新阶段 —— 数据化建设阶段，企业聚焦点逐步转向如何将企业内外部产生的数据高效应用，从而让企业决策不再依赖“拍脑袋”，而是靠“数据驱动”。

如今越来越多的公司有数据采集需求，如果说大数据是“屠龙术”，那么“龙”会越来越多。Gartner 预测，到 2020 年大数据将成为主流的嵌入式技术，并被视为常规产品的一部分。行业领导者（如互联网、金融、零售、企业级服务等）一直积极应用海量数据的采集和分析，聪明的初创企业都在拥抱数据驱动方法。

很幸运，我是“吃过猪肉，也养过猪”的人 —— 在百度的 8 年时间里，我见证并献身于百度大数据的建设，我所负责的团队从零到一构建了百度用户行为分析大数据平台，覆盖数据的采集、传输、建模、查询分析、数据可视化等前沿技术。

我注意到，我国绝大多数企业的数据化建设面临众多挑战，如数据采集缺失或埋点无序混乱、数据分析的能力欠缺等。2015 年 4 月，我从百度离职并创建神策数据（Sensors Data），目的就是将中国企业数据采集和建模等数据基础搭建好，让数据驱动真正在中国企业落地生根。

于是，我把本书命名为《数据驱动：从方法到实践》，希望我的思考与经历能够给读者一些启发。

第2章

大数据思维与数据驱动

我们先看一些国际上对大数据的应用情况。

苹果公司的传奇总裁史蒂夫·乔布斯在与癌症斗争的过程中采用了不同的方式，成为世界上第一个对自身所有 DNA 和肿瘤 DNA 进行排序的人。为此，他支付了高达几十万美元的费用。他得到的不是只有一系列标记的样本，而是包括整个基因密码的数据文档。

2008 年，Google 推出的一款预测流感的产品 —— 流感趋势（Google Flu Trends，GFT）。2009 年，Google 通过分析 5000 万条美国人最频繁检索的词汇，将其与美国疾病中心在2003 年到 2008 年间季节性流感传播时期的数据进行比较，并建立一个特定的数学模型。最终 Google 成功预测了 2009 年冬季流感的传播，甚至可以具体到特定的地区和州。

Prada 旗舰店中每件衣服上都有 RFID 码。每当顾客拿起任何一件 Prada 进试衣间，RFID 就会被自动识别。同时，数据会传至 Prada 总部。每一件衣服在哪个城市、哪个旗舰店、什么时间、被拿进试衣间后停留多长时间，数据都被存储起来加以分析。如果某一件衣服销量很低，以往的做法是直接下架。如果 RFID 传回的数据显示这件衣服虽然销量低，但进试衣间的次数多，那也许说明还有其他问题。

……

显然，数据能够给社会、企业带来商业模式上的优化，以及商业自动化的突破。拥有数据分析实践能力的企业正在进行数据资产的挖掘与利用。

究竟什么是大数据分析？它和传统的数据分析有何不同？

大数据的概念

在这一行业浸泡了近十年，不断有人问我什么是大数据？结合我看过的一些书籍，包括《大数据时代》、《数学之美（第二版）》、《硅谷之谜》，以及吴军老师的演讲等，我站在巨人的肩膀上，结合多年工作实践，形成了自己的一些认知。

我把大数据的概念总结为大、全、细、时。

大数据之“大”

2015 年，百度每天的行为数据超过 1.5 PB，我们毫不怀疑这是大数据。但全国各个地级市某一天苹果价格的数据大小只有 2 MB，相比前者可以忽略不计。如果我们基于这个数据，做一个苹果分销的智能调度系统，这就是前沿的大数据应用。Google 刚成立时，其创始人谢尔盖·布林和拉里·佩奇抓取了整个互联网的页面，虽然压缩后仅 47 GB，但是 Google 搜索显然是一个大数据的应用。而一台风机一天的振动数据大约 50 GB，但这个数据只是针对这一台风机的，因为覆盖面狭小，所以不能叫大数据。

因此，大数据的“大”强调宏观的“大”，而非一味追求数据量的“大”。

2014 年 8 月 24 日，美国南加利福尼亚发生了一场地震。传统的地震监测方法是在全国各地放一些地震监测设备，以此来监测地震的幅度。由于部署的监测设备较少，导致人们无法精确分析地震的影响。这次不同的是，随着运动手环的普及，分散在全国各地的手环，成了很好的振动采集设备。

图 2-1 是 Jawbone 手环公司在此次地震中收集到的振动数据，4 种线条代表了不同地区的数据，起伏趋势大致相同。地震发生时，最靠近震源的纳帕地区首当其冲，近 80% 的用户在睡梦中惊醒。该图与地震仪记录的地震时间、震源地区的数据非常相似。这就是一个典型的“大”数据应用。

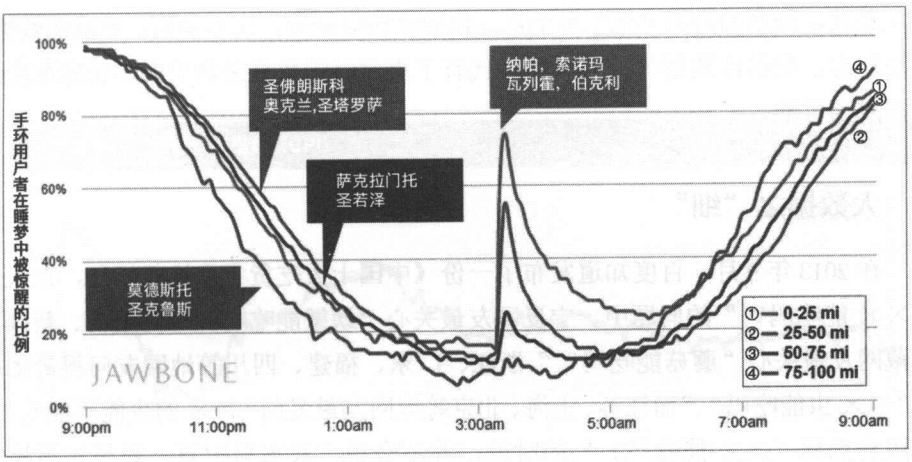


图 2-1　Jawbone 手环公司收集到的振动数据（图片来源于互联网）

大数据之“全”

我们再来看关于美国大选的三次事件。

1936 年《文学文摘》收集了 240 万份调查问卷，预测失败。

新闻学教授盖洛普收集了 5 万人的意见，成功预测罗斯福连任。

2012 年 Nate Silver[1] 通过互联网采集社交、新闻数据，成功预测大选结果。

《文学文摘》所收集的问卷有 240 万份，数据足够大，但为什么预测错误了呢？因为《文学文摘》是通过电话调查的，当时能够安装电话的只有富人，这类人群本身就有不同的政治倾向，调查结果本身就是偏的。盖洛普只收集了 5 万人的意见，但是他按照社会人群比例抽样，然后汇集总体结果，反而预测正确了。因为这次预测，盖洛普一炮而红，并成立了一家著名的调研公司。当然，后来盖洛普也有预测失败的时候。到了 2012 年，一个名不见经传的人物 Nate Silver 通过采集网上的社交、新闻数据做竞选预测。

预测的结果和真实的结果惊人地接近。当然，2016 年的大选，由于共和党的许多选民都是社会中下层，并不使用网络，导致这次所有基于网络数据的预测失败，Nate Silver 本人也不例外，可见网络本身也会带来有偏颇的数据。

[1] Nate Silver，号称美国公众眼里政治圈内完美的“预言帝”，他的“预言”被称为竞选预测之神谕。有媒体评论，其选情分析被极度精妙的美国政治评论圈认为达到了前所未有的水平，但因为他所使用的是被学界称为“巫术统计”的贝叶斯理论，也招惹到频率学派和一些保守的统计科学家们的质疑。

总之，我想强调的“全”是全量，强调多种数据源，包括前端、后端的数据，以及日志、数据库数据等。大数据时代有了更前沿的数据采集手段，让获取全量数据成为可能。

大数据之“细”

在2013年9月，百度知道发布了一份《中国十大吃货省市排行榜》，在关于“××能吃吗？”的问题中，宁夏网友最关心“螃蟹能吃吗？”内蒙古、新疆和西藏网友最关心“蘑菇能吃吗？”浙江、广东、福建、四川等地网友问得最多的是“××虫能吃吗？”而江苏、上海、北京等地网友最爱问“××的皮能不能吃？”当用户想问“××能吃吗？”的时候，并不会说“我来自宁夏，我想知道螃蟹能吃吗？”而是直接问“螃蟹能吃吗？”但是服务器可以采集用户IP地址，通过IP地址就能知道他们所在的省份，这就是多维度数据的威力。现有的采集手段，能够让我们从多个维度获取数据，在进行后续分析的时候，能对这些维度加以利用，这就是“细”。

总之，“细”强调多维度数据，包括事件、商品的各种维度、属性、字段等。比如我现在问“你所在公司的客户中，不同身高的群体在平均消费额上有什么差异”，你很可能回答不出来，因为你没有记录“身高”这一维度的数据。

大数据之“时”

CPI是居民消费者价格指数（Consumer Price Index）的简称，它是反映居民家庭一般购买的消费价格水平变动情况的宏观经济指标。

CPI是怎么统计的呢？其过程包括两个阶段：一个是收集商品价格数据，一个是分析并发布数据。我从百度百科上了解到，中国CPI采样涉及500多个市县，采价调查点有6.3万个，近4000名采价员，次月中旬发布报告。图2-2为2008年1月～2017年3月全国居民消费者价格指数情况。

美国有一家创业公司叫Premise Data[1]，它通过众包方式，由25000位采价员（学生、收银员、司机等），通过手机APP采集数据，每条6～40美分，比美国政府数据提前4～6周发布。如果企业或个人提前知道这些数据，就可以提前

[1] Premise Data，一个结合线上线下数据提供新鲜的经济观点的公司，比政府机构更准确地预测经济，并将有用的数据、指数和工具卖给有需要的公司。

做空或做多一些股票。据说 2008 年的金融危机，阿里巴巴就通过交易数据而更早有所感知。

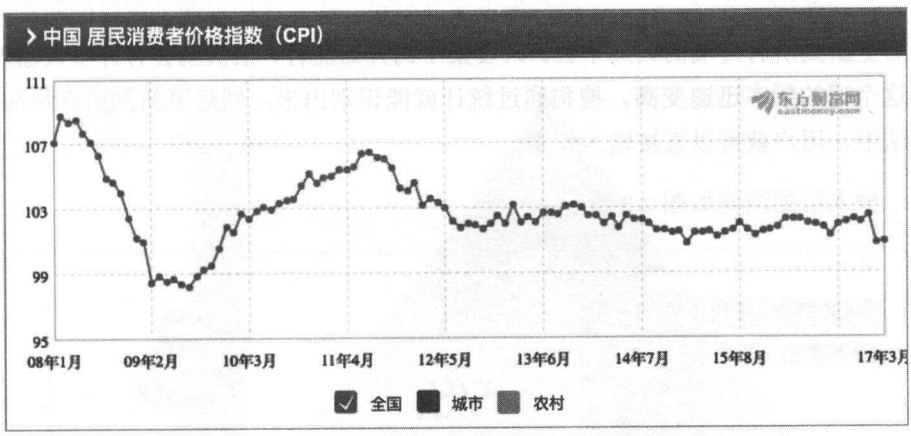


图 2-2　2008 年 1 月—2017 年 3 月，全国居民消费者价格指数情况（图片来源于网络）

这就是“时”，它强调实时数据采集和实时数据分析的价值。在 CPI 的例子中，我们可以让价格上报工作更智能，不需要人工的方式。

“大”“全”“细”“时”让我们对大数据的概念有较为清晰的认识，它们主要强调数据在获取和规模上与传统数据时代的差异。这是企业进行数据采集的“四字法则”，该法则对企业数据采集提出了一定的要求（详见第 3 章），企业有了夯实的数据基础，才能对大数据加以利用。

大数据的本质

运用大数据首先应该解决“思维”问题，大数据思维指的是企业在数据化运营和管理过程中运用数据的思维和方式。我们先来看两个案例。

案例 1：输入法的变革——从智能 ABC 到搜狗

智能 ABC 是一款古老的输入法，打字很慢，每次输入完毕，还要手工翻页选词，非常麻烦。2002 年左右，紫光输入法出现，当时带给我很大震撼，因为它采用了更好的词库和选词算法，让输入效率大大提升。使用时感觉按键还没按下去，字就已经跳出来了。但是其词汇更新比较滞后，往往半年才能更新一次，许多新词打不出来。

2006 年出现了一款云输入法——搜狗输入法，它直接基于搜狗的用户搜索记录和各种海量词库，词汇识别率高，词库更新快。用户平时的搜索和打字结果都会上传到搜狗服务器，基于这些数据做统计分析，就能直接识别出最可能的新词，然后更新到所有终端的词库中去。只要某个词开始流行，很快就会有许多人输入，让这个词的频率迅速变高。搜狗通过统计就能识别出来，然后更新到所有终端输入法中，用户就可以直接输入识别。

输入法的演变如图 2-3 所示。

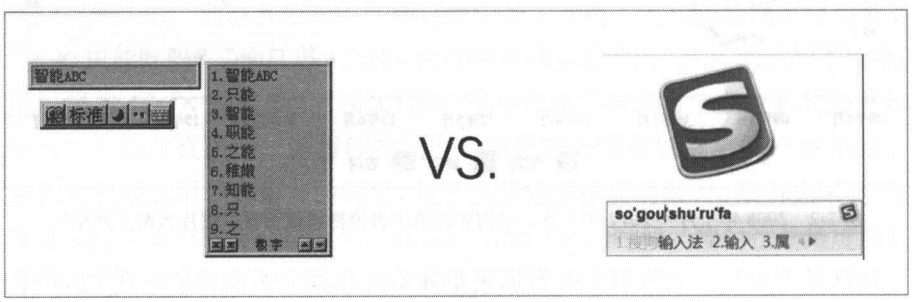


图 2-3　输入法的演变

案例 2：从纸质地图到百度地图的演变

地图已经存在了几千年，它讲述了人类永无止境的发现与探索旅程。如今地图不再是一张纸，人们会选择手机地图软件，如百度地图。地图软件承载了行政区、地点信息、道路名称和检索结果，原理是抓取用户的 GPS 信息分析人群流向与聚集情况，并从交管所等机构购买地面路况监测数据，从而对整个路况进行综合判断。同时以用户最好理解的方式搭建交互架构，每个层级上的信息都可以不断刷新，用户可以实时寻找地点、避开交通拥堵。

以上两个案例都通过数据分析与处理，带给用户截然不同的体验。随着各种前沿技术的发展，我们的思维方式已经从最直接的决策方式 —— 拍脑袋、因果驱动转化为数据驱动。直接向数据要答案，这就是大数据思维。我们获取的数据越全面，就越能消除更多的不确定性。

“大数据的本质是消除不确定性”，我第一次接触这个观点是在吴军的《硅谷之谜》一书中，当时觉得醍醐灌顶，我一直在思考究竟什么是大数据？而吴军的这句话直中要害。之后我在看《暗时间》一书时，尝试搜索信息论和不确定性的关系，发现克劳德 • 香农（Claude Shannon）说了这么一句话：“信息是用来

消除不确定性的东西。”果然是信息论的鼻祖，一句话解释了“信息”的精髓。

我们可以把信息分成 4 个层次，如图 2-4 所示。

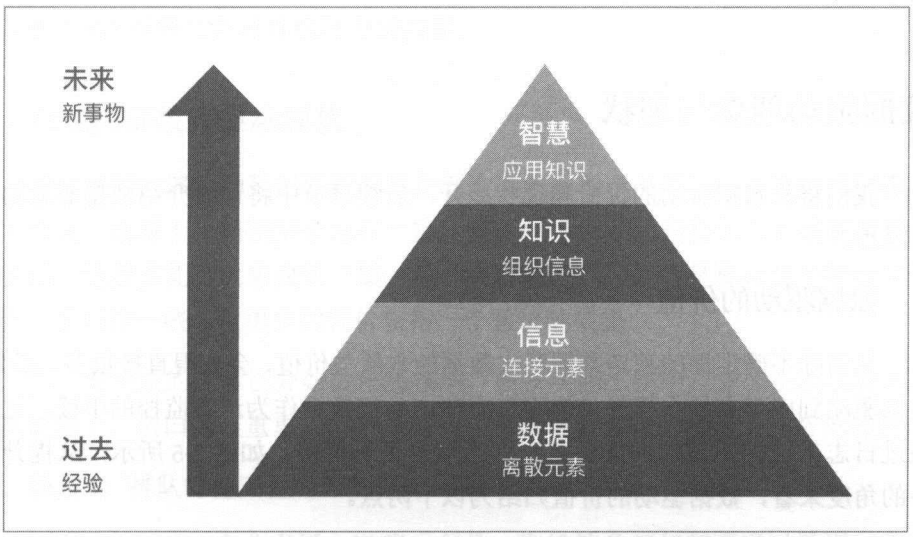


图 2-4　信息的 4 个层次

我们常说的数据，是信息的一种存储落地形式。比如你和朋友在交谈时，双方发生了信息传递，但是我们并没有把这些信息记录下来，也就没有形成数据。如果我们通过录音将信息录制下来，就形成了数据。数据是一类信息，而大数据又是一类数据。既然信息是消除不确定性的东西，大数据从本质上来说也是消除不确定性。

那么，什么是不确定性？

我们以天气预测作为一个例子。假如我现在让你预测某天某地的天气如何，这个时候你不掌握任何信息，只能像抛硬币一样进行猜测，也就是说你预测对的可能性是 50%。但如果我告诉你前一天是晴天，那么结果是晴天的可能性就大一些。如果我再告诉你大气云层、空气湿度、气温、风速等情况，你就能更加准确地做出预测。在这个过程中，当你掌握了更多的信息，也就消除了更多的不确定性。

再比如前面我们讲到的地图的案例。回家路上道路是不是拥堵？打开百度地图查看实时路况，就知道了答案。百度地图给你提供了信息，从而消除了这种不确定性。网站页面用蓝色背景好，还是绿色背景好？我们可以去做 A/B 测试，分

析哪种背景的点击率会更高。这与百度的企业文化之一——“用数据说话”是一个道理。数据有时候也会欺骗人，但大部分时候它还是客观冷静的，不带有感情色彩。

数据驱动理念与现状

我们将从数据驱动的价值和现状展开，后续章节中将详细介绍数据驱动的整个过程。

数据驱动的价值

从消除不确定性的视角来解释大数据的本质与价值，会变得直接很多。那么，数据驱动到底都有什么样的价值呢？有些产品把数据作为运维监控的手段，比如通过日志来监控系统的性能负荷，这当然也很有价值。如图 2-5 所示，从提升业务的角度来看，数据驱动的价值归结为以下两点。

其一是驱动决策。通过数据来帮助拍板，包括产品改进、运营优化、营销分析和商业决策等。我们有了数据，就能判断哪些渠道转化的效果更好，哪些功能样式更加受用户欢迎。这也就是我们常说的 BI（Business Intelligence，商业智能），通过数据来支持决策。

其二是驱动产品智能。所谓智能，我把它归结为这么一种模式：我们有了一定的数据基础，然后在上面套一个算法模型，再将得到的数据结果反馈到产品中。这样，产品本身就具有了学习能力，可以不断迭代。比如个性化推荐，通过采集许多用户行为数据，在这个基础上训练用户兴趣模型，然后给用户推荐信息，再将用户的使用数据反馈到模型中，精准广告就是类似的模式。智能是一种学习能力，产品智能就是现在比较火的 AI（Artificial Intelligence，人工智能）概念。

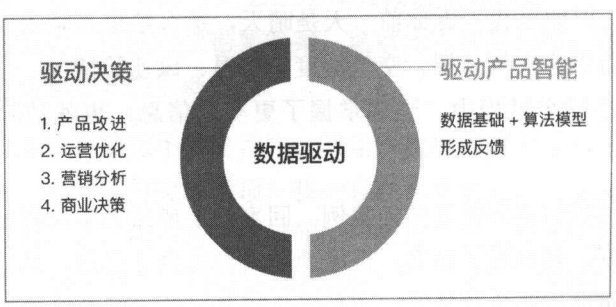


图 2-5　数据驱动的两大价值

这两点都消除了决策的不确定性，只是前者是人来执行决策，后者是机器来执行决策。事实上，我认为，数据驱动决策只能发挥数据 20% 的价值，甚至更少。而数据驱动产品智能将会发挥数据更大的价值，我也非常看好 AI 的发展趋势。第 4 章和第 5 章将重点讲解这两方面内容。

企业内部数据驱动现状

数据固然能够帮助我们看透笼罩在创造新业务和产品周围的不确定性阴霾，不可否认，这对于一些初创企业有一定困难：一个创业公司创始人无法拿到更多的数据，他需要凭直觉来决策“做一款什么样的产品”。但是要让这个阶段尽量缩短，更可控一些，以更少的代价获得一个验证的效果。

当一家企业的产品已开始被市场接纳，而在实际工作中，企业在实现数据驱动的道路上，依旧困难重重。以下是创业公司实现数据驱动道路上的常见场景。

场景 1：排队等待工程师跑数据

如图 2-6 所示，企业老板、运营、产品、市场等各部门都要通过数据工程师老王获取数据，整个流程包括沟通需求 → 分析数据源 → 升级数据采集系统 → 开发程序 → 提供结果等，老王忙得痛不欲生。当然，数据需求方都对数据获取的速度很不满意，有的人等不及，还是决定拍脑袋，最终导致产品迭代效率低下。

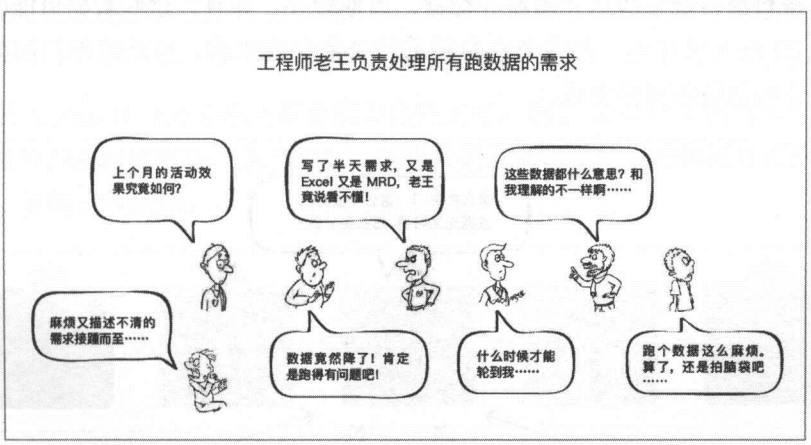


图 2-6　排队等待工程师跑数据现状

场景 2：仪表盘只能看到宏观数据

如图 2-7 所示，仪表盘能够帮助各个团队负责人看到宏观数据，如销售额、

用户数等，这在一定程度上帮助管理者做出科学决策。然而宏观的数据价值有限，这令执行者苦恼不已。比如昨天活跃用户数暴跌 20%，是什么原因？宏观的数据这时显然丧失价值，我们需要进行深入、精细化的分析，如按照渠道、地域等维度对数据进行分解，判断某渠道或某地域是否有大的波动，进行多维度、细粒度的下钻分析，才能快速定位问题，从而有的放矢地解决问题。

图 2-7　公司有仪表盘后的现状

场景 3：无法跨越数据孤岛的藩篱

如图 2-8 所示，企业内部的数据孤岛现象是普遍存在的，特别对一些集团化的企业孤岛效应更是明显。做大数据分析需要与不同部门沟通协调，获得审批权限，等待数据审批完成后才能统计数据，周期较长。并且，这些数据可能因为没有统一 ID 而无法打通。从企业自身数据的价值角度来说，应消除部门间的数据孤岛，让数据协作更好完成。

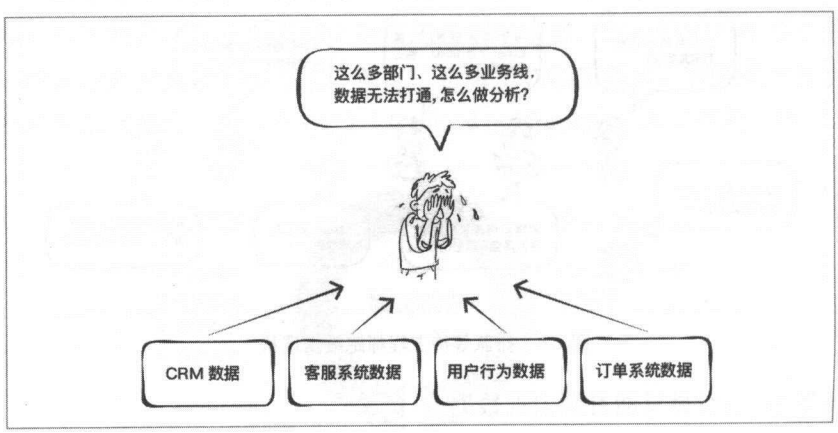


图 2-8　数据孤岛

理想的数据驱动——“流”

上述三个场景是典型的“需求驱动”，即根据需求去找数据。业务方提出数据需求，工程师满足需求，加上排队等待，整个效率非常低，完成一个需求都要几天甚至几周的时间。那么，理想的数据驱动应该是怎样的？

我们应该反向思考这一问题，先把数据源整好，在这个基础上提供强大的分析平台，让业务需求提出者能够自助式（Self-Service）地完成数据分析需求，从串行变成并行，完成需求从几天时间缩短到几分钟甚至几秒钟，这才是理想中的数据驱动，如图 2-9 所示。

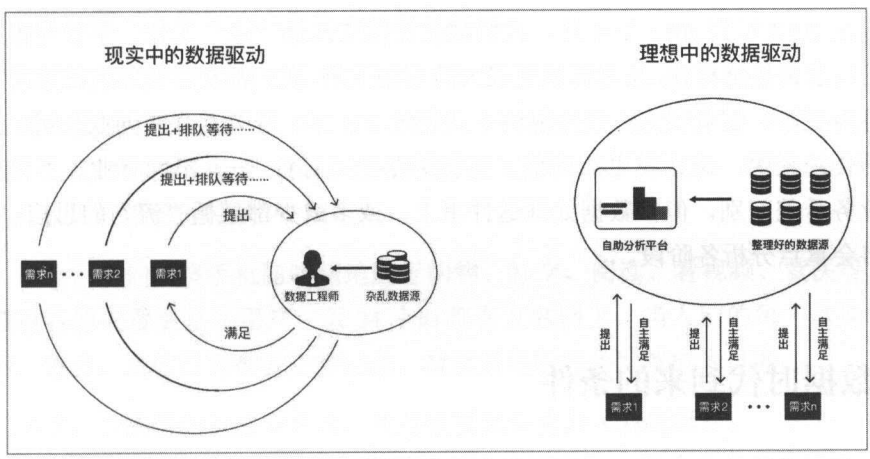


图 2-9　现实中和理想中的数据驱动

我从 2008 年开始专职从事数据方面的工作，到了 2012 年才慢慢想清楚——数据处理归根到底就是一条“流”。按照数据的流向，可以把数据处理分成 5 个阶段，如图 2-10 所示。

图 2-10　数据驱动的“流”

在这个过程中，每个业务人员和数据之间都需要有一个强大的工具，将数据规范化，处理数据模型。通过这个强大的分析工具，让这些业务人员在数据分析

平台上自助式地完成自己的分析需求，如图 2-11 所示。

图 2-11　自助式产品分析是最理想的数据分析方式

从 2012 年到 2015 年 4 月，我们都是在围绕这条“流”工作。不管是推进公司的日志采集结构化，还是提供更强大的查询引擎，我们都在尝试如何把这条“流”建设得更好。创业以来，我接触的企业超过 200 家，既有互联网创业公司，又有大的传统集团，这让我更加确信了这套思路的可行性 —— 不同行业、不同企业的业务千差万别，但在数据处理这件事上，或多或少都遵循“流”的思想。第 3 章将会重点分析各阶段。

大数据时代到来的条件

我们先看几个典故。王安石在变法过程中，向农户提供政府低息借贷，在不增加税收的情况下就提升了政府营收。庞涓在追孙膑的部队时沿途分析孙膑部队留下的土灶数据，土灶的数量呈递减趋势，并由此判断齐军士兵多数叛逃，便只带领少量精锐穷追不舍。然而这实际却是孙膑故意制造的“数据陷阱”。楚汉争霸，楚怀王与刘邦、项羽约定“先入定关中者王之”，刘邦先入，理应为王。项羽后入，烧杀抢掠。萧何建议刘邦找项羽说情，拿下不起眼的汉中。其实刘邦进入咸阳后，萧何直接把秦国的典籍、图册等资料收入囊中，经过深入研究发现，汉中税收占据秦国税收很大的比例。后来刘邦凭借汉中之地，打到三秦，并最终打败项羽。

可见数据思维与数据分析应用自古就有，而“大数据”的概念在 2011 年才火起来，在维克托·迈尔·舍恩伯格（Viktor Mayer-Schönberger）及肯尼斯·库克耶（Kenneth Cukier）编写的《大数据时代》一书中进行了详细描述。这一新技术与新趋势提出后，顿时引发行业追捧与热议却一直鲜有成效。而经过近几年的发展，泡沫逐渐退去，随着越来越多的数据驱动的应用落地企业，才进入了真

正意义的大数据时代。

接下来我们分别从数据采集、数据处理、数据认知三个方面分别介绍。

数据采集能力增强

数据采集能力是一个大数据团队必须具备的能力。根据前文提及的“大”、“全”、“细”、“时”，数据采集能力应该是全域数据的采集能力，包括 PC 互联网、移动互联网、IOT、线下数据等各个方面。大数据采集能力的增强原因有以下两点。

首先，移动互联网为大数据发展提供机遇。

移动互联网时代的到来重塑并颠覆了传统行业。据 Analysis 易观报告《中国互联网发展趋势报告 2016》指出，2016 年，移动应用快速增长，在社交、视频、新闻、工具和购物等领域，移动应用渗透率已超过 50%，在团购、旅游和零售业，移动端收入规模已经超越 PC 端。

如今，人们围绕手机随时随地进行购物、社交、阅读、看视频、就医等，可穿戴设备的火爆更是让用户一天 24 小时都在互联网上。当人们的每一次操作的时间、地点、具体行为都被完整记录，数据采集的完备性就成为可能。

其次，传感器的发展与普及，使得数据采集能力大幅度提升。

谈到传感器，我们不难想到美国亚马逊推出的“亚马逊 Go”超市，这是一种无须结账的新商店。店中装有利用机器学习和算法的传感器，顾客不用排队便可自动结账。

图 2-12　“亚马逊 Go”顾客可以拿走想要的商品，随后直接离开

在我国，传感器的发展与普及正在支撑我国智慧城市、智慧交通、智慧能源、智慧医疗、智慧环保等的建设。创业公司蜗牛睡眠和 DFocus 在传感器应用方面颇为前沿。

蜗牛睡眠 APP 通过人体呼吸律动、体动的幅度来影响传感器，以此记录用户的睡眠数据，测试每天深睡、浅睡及醒着的时间，并生成一张柱状统计图，然后给出睡眠质量的打分，用户可以非常直观地了解自己的睡眠状态。

DFocus 利用物联网传感器、数字终端和大数据分析等，帮助企业有效降低工作场所的成本，并提高空间使用率。智能监控平台监控能源和空间使用情况、移动侦测状况及温度和环境信息，以便能更好地观测建筑空间的使用情况。这些数据被存储下来，可供图形审查、详细分析或建筑管理系统使用。

传感器作为物联网中一个从外界接收信息的载体，被誉为物联网、智能设备的“心脏”。我认为，大数据的未来就是传感器的时代，这也是我将公司命名为“Sensors Data”的原因。

线下数据采集分析是时下零售业掘金大数据的热门应用，传感器的普及与发展提升了数据采集能力。零售商希望能拥有类似在线电子商务网站的 Cookie 一样记录顾客的行为模式、偏好和转化率等数据的工具。他们希望了解有多少客户看了 A 产品？有哪些客户触摸过 B 产品而放弃购买？有哪些客户是回头客？这些顾客之前买过哪些产品？这些数据的采集依赖各类型传感器的应用，如手势传感器、手指弯曲传感器、动作追踪传感器、触觉传感器、眼动追踪等，能够实时跟踪用户行为，并分析出每一位进店客户的交易历史、行为举止、兴趣偏向等决策信息。

数据处理能力增强

数据处理能力是对数据的采集、存储、检索、加工、变换和传输的能力。数据处理是实现数据分析和挖掘数据价值的前提，是衡量大数据发展状态的重要指标。摩尔定律及大数据分析和计算技术的发展，赋予了企业很强的数据处理能力。

摩尔定律以更低的成本获得更高的处理能力，揭示了信息技术进步的速度。在大数据时代，随着时间的推移，我们能够获得的硬件、软件处理能力不断增强，硬件费用、软件费用、耗电量等运营维护成本也在不断降低。数据计算和数据存储成本一直在下跌，让大数据相关技术的覆盖更为广泛。摩尔定律催生的更快、更高效的大数据技术带来了巨大的经济效益，让大数据引发的企业数字革命成为现实。

大数据技术中大数据的分析和计算是核心。传统数据处理技术的瓶颈在于分布式文件存储和并行计算能力，海量的存储和计算在一定程度上制约着企业的数据应用。如今，大数据分析和计算的新技术在飞速发展与迭代中。

当我在 2008 年刚接触 Hadoop 时，Hadoop 的成熟度还很低。一名工程师可能得花两周时间才能让它运行，系统能够支撑的最大机器数不过几百台。而现在，一名工程师只要下载一套 Cloudera 社区的 Hadoop 版本，两个小时就能运转起来，并且最大的集群规模可以达到几万台。Spark 是在 2011 年兴起的另一套大数据平台，在支持机器学习类迭代计算上，有更好的性能。现在 Hadoop 和 Spark 生态可谓非常成熟。

总之，技术的成熟大大提升了数据处理能力。

数据意识的提升

《大数据时代》一书提及大数据开启了一个重大的时代转型。就像望远镜让我们感受宇宙，显微镜让我们能够观测到微生物一样，大数据正在改变我们的生活以及理解世界的方式，成为新发明和新服务的源泉，而更多的改变正在蓄势待发。

大数据时代正在引发深刻的思维转变，大数据改变每个人的日常生活和工作方式，改变商业组织和社会组织的运行方式。

大数据概念出现时伴随着泡沫炒作与热情追捧。新技术一出现，人们甚至寄希望于其可以改变世界。尘埃落定后，记录、存储和分析等方面冲破了技术限制，人们逐渐意识到拥有了可以采集、处理大规模数据的能力。同时，丰富的数据分析与商业智能实战项目落地，数据的应用频率不断上升，引发企业对数据资产的重视，如何让数据服务企业成为企业的思考点。

互联网作为新兴行业，在高速发展初期引发诸多流量红利。如今，红利已经消失殆尽，企业认识到，必须通过数据分析去了解用户需求、洞察用户心理，从而提升用户体验，最终构建起自身的核心竞争力。企业发展已进入精细化运营阶段，建立起以用户为中心的设计、数据驱动的产品管理意识。5 年前，大家还在讨论数据重不重要的问题，现在问题的焦点已经变成如何让数据发挥更大价值。

大数据时代已来，数据分析让一些商业成功有迹可循。

第 3 章

数据驱动的环节

我把中国的企业信息化建设分为两个阶段：第一个阶段是从 2000 年到 2015 年，主要是企业纷纷选型信息系统，包括财务、ERP、CRM、官方网站等。这个阶段以用友、金蝶为代表；第二个阶段是从 2015 年到 2030 年，有了前 15 年的基础，企业开始重视数据，走向数据化建设，兴起了一批 SaaS 服务企业，包括数据分析、数据可视化相关的创业公司。但如果对比美国的信息化建设历程，你会发现它们比国内提前 10 年甚至更久。数据驱动的理念在 2000 年左右就已经开始在美国普及，那些企业也有更好的 IT 基础。

而在国内的企业中，BAT 最早推进数据化建设，可以说从公司成立之初，就有了数据的基因。不管是百度的搜索引擎，还是阿里巴巴的电子商务，无不围绕数据进行服务。百度的理念是技术驱动，在大数据处理方面，格外舍得下功夫。

上一章提到，在 2012 年，也就是我从事数据工作三年多后，我逐步认识到数据处理归根到底是一条“流”，这条“流”包括数据的采集、传输、建模存储、查询分析和可视化。所谓数据建设，就是不断地完善这条流。就拿数据采集来说，从非结构化的文本日志，到结构化的数据源，或者从数据传输来看，从批量离线传输，到实时传输。这些工作都是在让整个系统变得更好。当然，这是从技术架构的角度看待数据驱动。

如果单纯从数据的角度来看，我们可以把它分成 4 个环节：数据采集、数据建模、数据分析和数据指标，如图 3-1 所示。这个认识也是在我创业这两年才逐步清晰起来的。

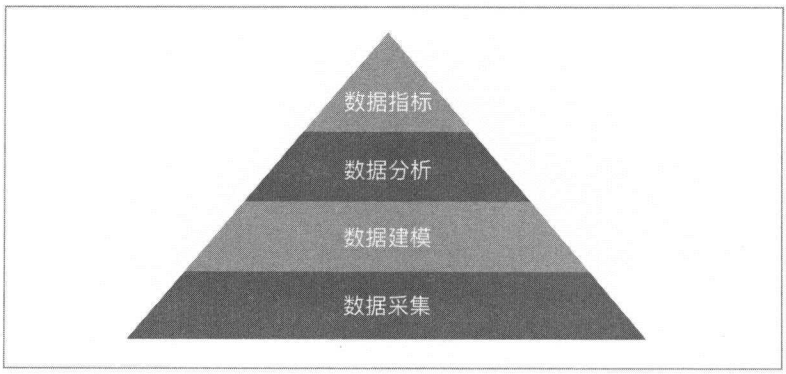


图 3-1　数据分析金字塔

数据采集与埋点

提到数据分析，许多人会问有没有特别有效的分析方法，让业务出现大的飞跃。这就像在战争中，只考虑有没有出奇制胜的打法，却不知真正的战争高手，往往是连粮草供给都能管理好的。这也是孙子在《孙子兵法》中所说的“不败而后求胜”。在航天事业中，大家往往关心卫星多么高级，但真正的挑战恰恰是送卫星上天的火箭技术，这就是埃隆·马斯克的火箭回收技术引人注目的原因。

同样，在数据分析的整个体系中，数据采集恰恰是最重要的。我在第 1 章提到，数据源很重要，这是我在百度做大数据时的最大心得。数据源和数据建模，恰恰是冰山下面的部分，各种分析方法只是冰山一角。我们建设好数据源，就做好了数据的根基。

在数据采集上，“埋点”一词非常形象，我是在创业之后才知道这个词的。因为我在百度时的数据采集是基于日志的，而非在业务逻辑上嵌入代码，在正常的业务逻辑中嵌入数据采集代码的过程，就是“埋点”的过程。

数据采集的现状

企业在数据采集的道路上经常会遇到各种各样的问题，充斥着痛苦、幻想和绝望。

困惑：如何采、采哪些、用什么手段

一般创业公司的数据采集工作，通常会选择三种途径，分别是第三方统计工具、通过业务数据库做统计分析和 Web 日志统计分析。

其中，友盟、百度统计等第三方统计工具，通过嵌入 APP SDK 或 JS SDK 来直接查看统计数据。这种方式简单、免费，基本满足宏观基础数据分析需求，如访问量、活跃用户量等。但这类统计工具的用户很快发现了三个问题。

1. 由于数据采集不完整，无法实现深度分析。这种方式的 SDK 只能采集到一些基本的用户行为数据，如设备的基本信息、用户执行的基本操作等数据，而服务端和数据库中的数据并没有采集。即使是客户端的数据，也无法采集到一些精细化的维度。例如，在一些提交操作中，提交订单对应的成本价格、折扣情况等信息无法采集，导致后续的分析成了“巧妇难为无米之炊”。

2. 统计不准，与业务数据库对不上，甚至丢数据。这是前端数据采集的先天缺陷，后续将详细介绍，网络异常、统计口径不一致等因素，都会导致数据对不上。

3. 云模式的数据分析平台让不少企业有安全顾虑，不愿意将核心数据放在第三方平台上。

通过业务数据库实现统计分析时，一些互联网公司基于业务数据库中存储的订单、用户注册信息等数据，进行常规的统计分析需求，实时且准确，但也有不足之处。

首先，业务数据和统计分析数据耦合。业务数据库是为业务运转而设计的，满足机器读写访问需求。为了提升性能，会进行一些分表等操作。一个正常的业务都要有几十张甚至上百张数据表，这些表之间有复杂的依赖关系，这就导致业务分析人员很难理解表的含义。运营人员硬着头皮用几个月的时间好不容易看明白了，可能隔天又被工程师告知因为性能问题拆表，导致运营人员做无用功。

其次，性能较差，无法进行批量数据操作。业务数据表设计针对高并发、低延迟的小操作，而数据分析常常针对大数据进行批量操作，导致性能很差。

最后，缺少必要的数据字段。业务数据库是为满足正常的业务运转服务的，而有些分析需求用到的信息并不会在业务数据库中出现。比如浏览器版本信息，我们在进行数据分析时就会用到，分析不同浏览器版本的用户转化情况，但是正常的业务流程并不使用，这时我们就无法进行对应的分析。

使用 Web 日志统计分析，即用户在进行各种访问时，在服务器端打印一条记录，这条记录包含本次访问相关的信息。该方法能实现数据的解耦，使业务数

据和统计分析数据相互分离。然而，这种方式的问题“目的不纯”——Web 日志往往是工程师为了方便 Debug 顺便做的，这样的日志对于业务层面的分析，常常“缺斤少两”。另外，从打印日志到处理日志再到输出结果，整个过程很容易出错，我在百度时花了几年时间才解决了这一问题。

不可否认，以上三种方式都一定程度上解决了一部分数据采集的问题，但并不彻底。

痛苦：埋点混乱，常现埋错、漏埋

我曾经接触了一家七八年的老牌互联网公司，他们的数据采集有 400 +个点。每次数据产品经理 A 提出数据采集的需求后，工程师 B 就会按照要求增加埋点，并交给数据产品经理 A 去验证。A 最初觉察不到异常，但是产品上线之后，却发现埋错了，或者漏埋了，然后要求 B 再进行升级发版操作，整个过程效率极低。这是不少企业埋点的缩影。

无奈：数据团队和业务工程团队配合困难

一般来说，企业 A 轮融资之后，会有专门的数据团队或者兼职数据人员来负责企业的业务指标。为了拿到基本的业务指标，需要业务工程团队去配合做一些数据采集工作。在两个团队配合方面，以下两种原因让数据采集工作不能得到应有的重视。

首先，求“快”，数据分析让路产品升级。

产品迭代通常是企业优先级最高事项，当数据采集工作与产品迭代撞车时，一般会放弃数据采集工作。如果没有数据指标的支撑，就无法衡量这个功能的升级是否合理。互联网产品并非功能越多越好，产品是否经得起考验，还是要靠数据说话。

其次，KPI 驱动，数据团队需求得不到业务重视。

数据团队和业务工程团队是平级的团队，数据团队工作烦琐且不能直接提升工程团队的 KPI，导致需求时常不被重视，让数据采集工作难有进展。

数据采集遵循法则

欲流之远者，必浚其泉源。让数据驱动落地企业，数据采集的质量将决定数据分析的深度。其中，数据源是最重要的。一个查询引擎的好坏，无非是查询时

间的消耗差异，都会拿到正确的结果，但速度会快 20%，有时候影响并没那么大。若数据源存在问题，那么无论数据处理应用多智能的算法，都无法得出正确的结论。

关于用户行为数据采集，在核心逻辑里面，企业要将前后端记录的行为事件、关键维度信息记录下来，例如与交易相关的核心数据信息与维度信息都应该被记录。另外，关于人工审核（申请材料、抵押物审核等）的一些数据，我们也应该实时、批量引入。这与第 2 章提到的大数据“大”“全”“细”“时”概念如出一辙。

这四字法则对企业数据采集提出了哪些要求？

“大”强调的是宏观的大。这不只需要海量数据，还要从系统的角度去考虑。比如一个二手车买卖网站，我们要考虑买车方、卖车方、经纪人、检测员等多种角色。

“全”强调多种数据源。对于用户行为分析来说，不但要采集客户端数据，还要采集服务端日志、业务数据库，以及第三方服务器等数据，全面覆盖。而且，企业要采集全量数据，而不是抽样，如图 3-2 所示。这是和传统统计分析的一个显著区别。比如，我们不能抽取了个别省份的数据，就开始进行全国分析，有些省份会存在特殊性，可能导致错误的结果。

图 3-2　遵循“大”“全”“细”“时”的原则，数据采集应包括前后端数据、业务数据等

“细”要求把不同的维度都采集下来。对于用户行为事件来说，我们要采集 Who、When、Where、How、What 等信息，完成每一个事件的维度 / 属性 / 字段的采集，对事件的发生形成一个快照。若数据采集只是按照需求采集，在后续数据分析时不断遇到新需求，则又要采集新的数据，如此会延长整个迭代周期，导致效率低下。

“时”则强调时效性。比如“双 11”活动，有海量的用户涌入进来，我们就

要实时调整策略，比如把购买情况不好的产品替换掉。如果你不能实时查看数据情况，可能就会错失良机。再如在消费金融领域，我们要根据用户申请等行为特征，进行反欺诈判断，这些信息的采集必须是实时的。企业只有实时采集和分析数据，才能保证分析结果的最大价值。

在实际工作中，“大”可以作为宏观的考虑原则，重点关注更“全”和更“细”，而“时”可以根据业务场景灵活把控，毕竟数据的时效性是有成本的。

科学的数据采集和埋点方式

本节将介绍如何进行科学的数据采集。总的来说，数据采集方式归结为可视化 / 全埋点、代码埋点和导入辅助工具三类。

可视化 / 全埋点

2010 年，百度 MP3 团队做了一个叫作 Click Monkey 的产品，只要页面上嵌入 SDK，就可以采集页面上所有的点击行为，并可以绘制出用户点击的热力图，这种方式对于一些探索式的调研还是非常有用的。到 2013 年，国外一家数据分析公司 Heap Analytics，将这种方式更近一步，将 APP 的用户行为尽可能地全面采集，然后通过界面配置的方式对关键行为进行定义，这样便完成了所谓的“无埋点”的数据采集。使用这种方案，必须在产品中嵌入 SDK，等于做了一个统一的埋点，因此“无埋点”的叫法实际上是“全埋点”的代名词。

无埋点具有以下优势。

1．可视化展示宏观指标，满足基本数据分析需求。通过展现 PV、 UV 等网站或 APP 分析的宏观指标，告诉运营人员每个控件被点击的量有多少，哪些控件值得做更进一步的分析，以助于企业了解用户行为，为进一步数据分析指明方向。

2．技术门槛低，使用与部署较简单。只需要嵌入 SDK，极大程度避免了因需求变更、埋点错误等原因导致重新埋点的复杂工作。

3．用户友好性强。可以直接应用手指或者鼠标进行操作，自动向服务器发送数据，避免手工埋点的失误。

作为前端埋点的方式之一，无埋点有先天缺陷，它带来易用性的同时，也牺牲部分数据的采集深度。无埋点的劣势如下。

1. 无埋点只能采集到用户交互数据，且适合标准化的采集，自定义属性的采集需要代码埋点来辅助。

每个用户的交互行为均有许多属性，无埋点无法深入到更细、更深的粒度。例如在电商行业中，用户点击“购物车”是一次交互行为，无埋点会忽略用户信息、商品品类等其他维度信息，此时需要配合代码埋点来辅助数据采集；再如用户上滑屏幕时，内容瀑布流的底部载入、商品或广告的加载展示、下拉菜单中下拉内容的数据点击等情况，这类自定义行为的采集需要代码埋点来辅助实现。

由于无埋点仅适合标准的方案采集，一些数据分析平台也开始支持用户为每个事件添加自定义属性，如此能大大扩展事件分析的效能。

2. 无埋点兼容性有限。在安卓系统进行埋点时，不同工程师可能会给 APP 界面中相同的 Button 起不同名称的 ID，当运营人员想筛选出所需数据时，不同名称会给运营人员带来困扰。另外，由于目前第三方框架较多，如 RN 框架，容易造成无埋点兼容性问题。

3. 无埋点是前端数据采集方式之一，因此具有前端埋点的天然缺陷，如数据采集不全面、传输时效性较差、数据可靠性无法保障等问题。无埋点的技术原理依赖网站或者 APP 后端技术开发的严谨性与规范性、网络状态、网络口径等因素。

图 3-3 是无埋点的优劣势分析汇总。

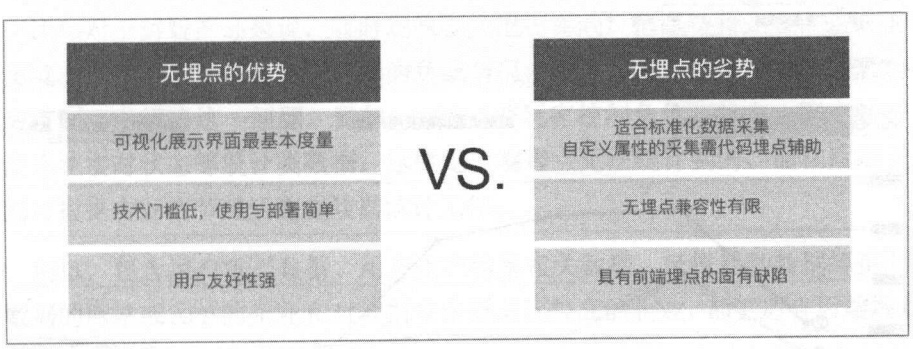


图 3-3　无埋点的优劣势分析

代码埋点

代码埋点又分为前端代码埋点和后端代码埋点。前端代码埋点类似于全埋点，都是在前端嵌入 SDK 的方式，所不同的是，对于每一个关键行为，我们都需要调用 SDK 代码，将必要的事件名、属性字段等写入代码，然后发送到后台数据服务

器。后端代码埋点则将相关的事件、属性等通过后端模块调用 SDK 的方式，发送到后台服务器。

这种方式相比全埋点来说，更适合精细化分析的场景。我们可以将各种细粒度的数据采集下来，方便做后续的深度分析需求。其中后端代码埋点，相比前端代码埋点，具有更高的数据可靠性，并且可以实现一处埋点，不用从各个 APP、Web 端进行埋点操作。

导入辅助工具

为了减少系统耦合性，我们还可以采用日志、数据库的方式生成数据，然后对数据进行转换，通过实时或批量工具完成数据导入。对于离线数据，比如线下人员和客户沟通情况等，可以通过导入工具完成数据采集。事实上，我在百度经常将日志格式的数据通过 LogAgent 模块实时传入后台服务器，也会采用分布式抓取的方式，定时将数据从源头下载到数据服务器上。

如何选择采集方式

面对这三种数据采集方式，我们该如何选择呢？要真正实现精细化运营，企业数据采集所采用的埋点方式不应“千企一面”，而应该“因企而异”。

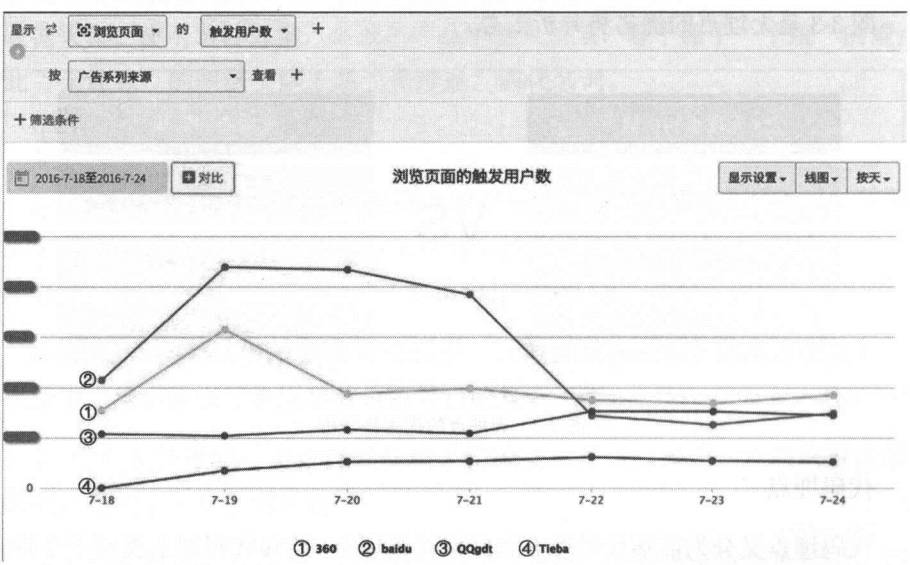


图 3-4　不同渠道和推广方式的效果分析

1．全埋点 VS. 代码埋点

如果仅仅为了看看宏观数据，并没有精细化分析需求，并且是对客户端做的分析，这种时候全埋点是一种比较省事的选择。如阅读类、词典类工具性 APP 的企业客户，在其发展初期的产品运营阶段，产品功能较为基础，无明确业务数据、交易数据，仅通过 UV、PV、点击量等基本指标分析即可满足需求。如果全埋点还采集了渠道来源信息，则可以进行不同的渠道来源对比。图 3-4 是某广告企业通过全埋点的方式采集数据后了解用户渠道来源，并判断不同渠道和不同推广方式的投放效果。

一旦企业有复杂的分析需求，就必须进行代码埋点，否则数据无法进行灵活下钻。

2．前端埋点 VS. 后端埋点

在产品运营的初期，产品功能比较简单，可以采用前端埋点。或者有些行为没有和后端进行交互操作，比如有些游戏是离线运行，就比较适合前端埋点。

为了保证核心数据的准确性，我们更推荐“后端埋点”。当前后端都可以实现数据采集时，应优先考虑后端（代码）埋点，尤其在各行业中有特殊业务需求的数据，更是强烈建议通过后端（代码）埋点方式采集。总的来说，后端（代码）埋点，或者“后端（代码）埋点 + 全埋点”方案，适合有深度数据分析需求的企业。

比如对于游戏产品来说，有时玩家已经退出游戏，但是链接还在，这时前端采集就不准，无法正确衡量服务器的负载情况、数据库的压力情况等，而后端代码埋点则可以解决这一问题。再如，NPC（非玩家控制角色）状态、副本状态、经济系统实时状态等统计类数据，这些是前端埋点无法统计到的，而在后端采集数据可根据实际情节灵活完成数据统计工作。

所以，包含用户资产数据、用户账户体系相关数据、风控辅助数据等重要业务数据的网站或 APP 的企业和对数据安全要求比较高的企业，都更适合后端埋点。

从后端采集数据，例如采集后端的日志，实质上是将数据采集的传输与加密交给了产品本身，认为产品本身的后端数据是可信的。而后端采集数据到分析系统中则是通过内网进行传输，这个阶段不存在安全和隐私性问题。同时，内网传输基本不会因为网络原因丢失数据，所以传输的数据可以非常真实地反映用户行为。

图 3-5 总结了适合前端全埋点和后端代码埋点的企业需求。

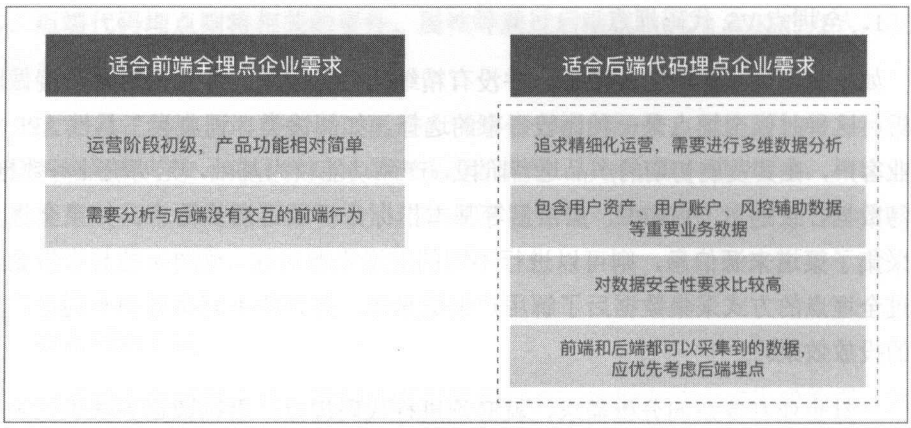


图 3-5　适合前端全埋点与后端代码埋点的企业需求

可见，数据源很重要，我们要更“全”更“细”地采集数据。

无论是 SDK 埋点还是后端实时或批量数据导入工具采集。这些都是手段，需要根据不同的应用场景，灵活设计数据采集方案。

接下来，我们探讨一下数据的准确性问题。

数据的准确性

在进行数据统计时，我们经常会对数据的准确性产生怀疑。如果出现较大偏差，比如 200% 的偏差，就很容易发现数据是不对的。但如果数据只有 5% 的偏差，就可能很难感受到。我们经常也会发现，用第三方统计工具所得到的数据和业务数据库中的数据对不上。

数据不准确的原因有以下几种情况。

1．网络异常

网络异常是导致数据不准确的直接原因之一。比如我们在使用 APP 时，可能因为网络异常，导致用户的操作行为并没有被及时发送到统计服务器端；或者这些服务是公共的 SaaS 服务，在一些网络的高峰期，比如晚上 8 点，同时有大批的 APP 用户向服务提供商发送行为数据，这样就容易导致网络拥堵，就像春运期间在 12306 网站抢购车票一样，容易导致某些请求丢失，造成数据不准；再者为了应对网络异常，我们通常会采用重传、间隔上传之类的策略，而这些策略由于标准不统一，也会带来统计的不一致。对于 APP 来说，在发送过程中，缓存到本地的数

据如果到达上限，可能会造成部分数据丢弃。JS SDK 通过同步发送，更容易出现丢失。

2．统计口径不同

同样是活跃用户应该如何统计？启动 APP 就属于活跃？还是首页加载完成才算活跃？用户是否要完成其他关键行为？这里就极易出现偏差。如何定义一个新用户？有多久没有再次应用产品才算新用户？另外由于用户 Cookie 被清除的处理策略不同等因素，看似同一指标，却会造成实际数值的偏差。

3．代码质量问题

一方面，由于众多的手机生产厂商及手机操作系统的发行版本，以及 APP 的开发框架和程序代码质量等问题，可能会导致 SDK 在某些情况下不能被有效调用，或者重复发送，这样也会导致数据的准确性问题。另一方面，负责埋点的工程师，可能因人为失误漏掉某些行为事件的采集。

4．无效请求

比如竞争对手的恶意攻击，Spider 等进行的抓取操作，都会导致数据的异常。

总的来说，我们并不能保证数据的绝对准确，毕竟代价太大。对于业务统计分析，没有必要像银行的转账系统那样具有高度的数据准确性，但我们可以采用一系列策略来提升数据准确性，让关键行为可以接近 100% 的准确率。

提升数据准确性的具体策略主要包括以下几个方面。

1．采集关键行为，推荐后端埋点。

对于关键行为，如订单交易、注册等，推荐后端 SDK 埋点或通过 LogAgent / Importer 进行数据采集。在前文数据采集方法中已经做过详细介绍，这里不再赘述。将关键行为通过后端数据采集，可以提升数据的准确性。

2．进行事件设计和明确统计口径，保证统计数据的质量。

统计口径不同，是人为因素中最大的问题。一般 SaaS 统计系统是采用通用的统计指标模型，即不管是什么 APP，均采用一样的统计策略。但对于客户的自有产品来说，总有一些自己的特殊定义方法，特别是针对一些边界用户[1]的处理。对此，我们还是要进行事件设计，以保证统计质量。

[1] 边界用户，例如一个用户在一段时间内，通过 EDM、沙龙活动两个渠道来到官网首页，在进行统计时，该用户应该被计入两个渠道，还是被计入其中一个渠道？这个用户则属于边界用户。

具体来说，分析师要和业务人员一起梳理业务流程，并拆分出关键用户行为事件（Event），以及事件相关的维度。在进行一些多维分析时，如漏斗分析、留存分析，以及基于事件与维度进行统计，数据会更加精确。比如，在嵌入多个子页面的一个页面中，加载过程应该算几次 PV？事件模型的内容后续会进行专门介绍。图 3-6 展示了 APP 浏览页面总次数。

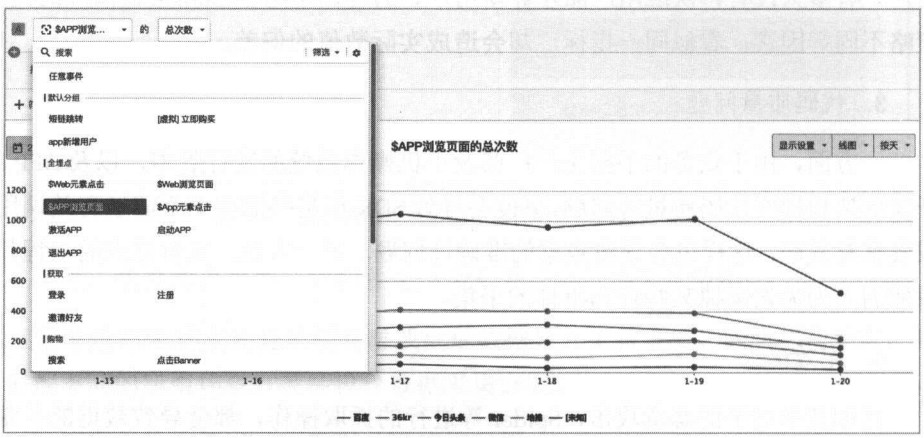


图 3-6　APP 浏览页面总次数

3．需要具有完善的元数据管理和埋点管理。

元数据管理可精细化管理每个行为事件的属性类型定义，埋点管理会跟踪每个埋点的数据量、校验通过量等，出现异常要及时发现。系统还应提供 Debug 模式，以精确跟踪每一条数据是不是按预期的计划进行处理。实时导入监测功能，可以通过用户 ID 及其他属性，筛选实时采集的数据，精确定位问题。总之，是要把数据采集从“黑盒”变“白盒”。图 3-7、图 3-8 展示了在神策分析平台实现的元数据管理和埋点管理，图 3-9 是神策分析实时导入监测功能。

4．通过多维分析能力快速定位异常。

分析工具要提供灵活的多维分析能力。这可以保证数据实时接入和实时查询，加上底层的数据是事件级的粒度，可以跟踪用户的每一步操作以及筛选相关的维度。这样在追查数据异常的时候，通过多维分析和用户详细轨迹，很容易快速地定位问题。我们曾经有个客户交易数据对不上，后通过多维分析，发现某种支付渠道的数据在系统里，而在已统计的平台中漏掉采集。图 3-10 是神策分析灵活多维分析之 APP 激活的人均次数。

通过以上一系列策略，我们就可以让数据采集达到较高的质量标准。

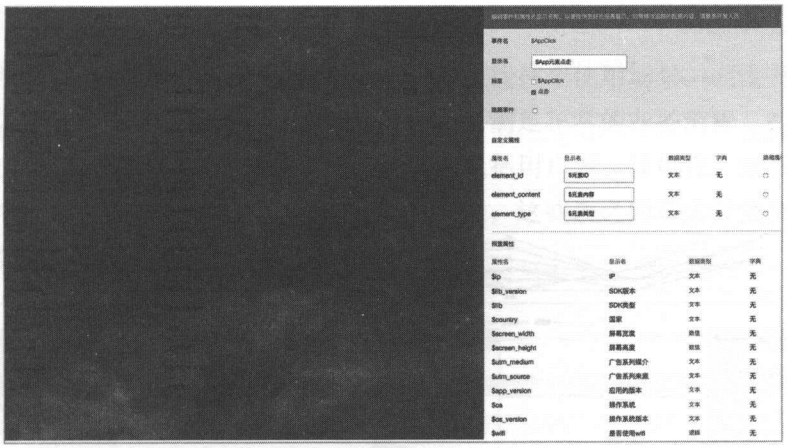


图 3-7 在神策分析平台完成元数据管理

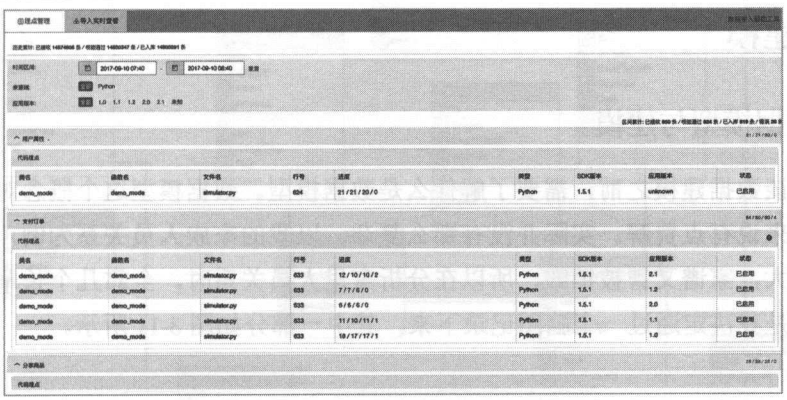


图 3-8 在神策分析平台完成埋点管理

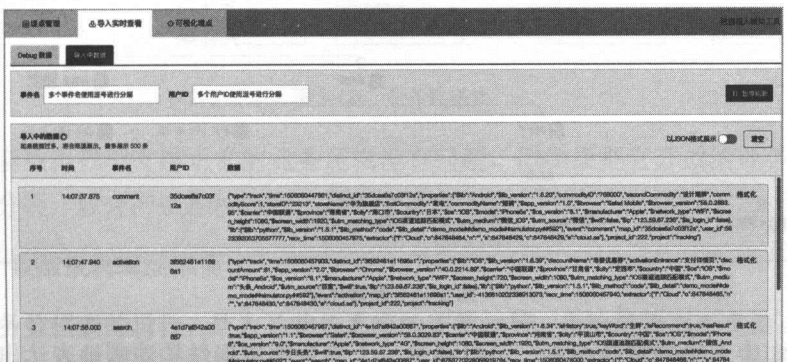


图 3-9 神策分析实时导入监测功能

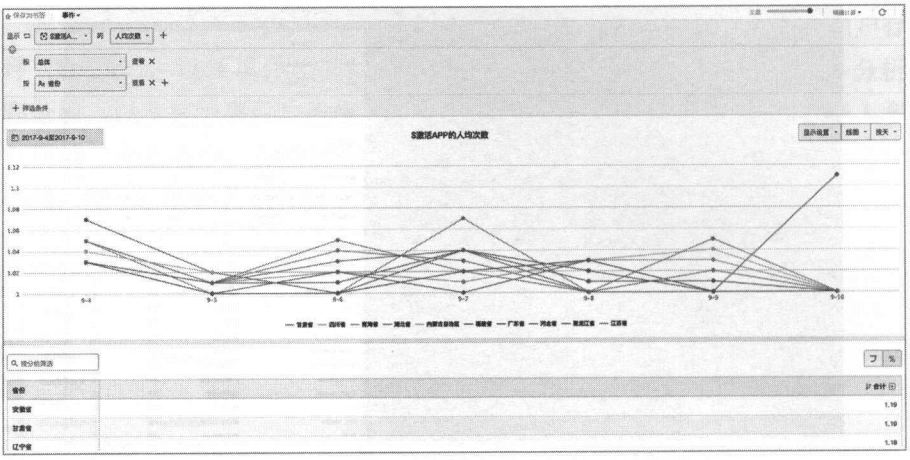


图 3-10　神策分析灵活多维分析之 APP 激活的人均次数

数据建模

数据模型与建模

在谈数据建模之前，需要了解什么是数据模型。数据模型这个概念对许多业务人员来说有点费解，实际并没有那么复杂。以我的家族人员关系为例，我的家族比较大，家谱又曾被销毁，所以在分析家里人员关系时，全靠几个长辈脑袋记忆。于是我决定通过一个脑图记录下来，其中一部分如图 3-11 所示。

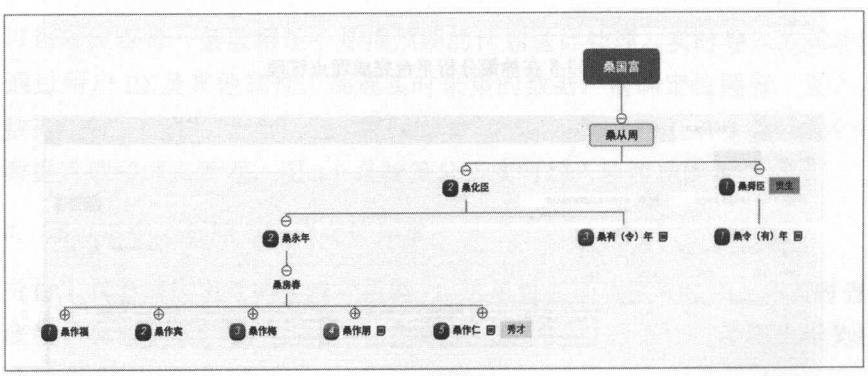


图 3-11　家谱片段

现实中人物的复杂关系，可以通过一个简单的树形结构清晰地表达出来，这就是一个数据模型。当然，我们也可以通过一个 Excel 表格，每行记录一个人名，再列出他的父亲是谁，兄弟是谁，这也是一个数据模型。简言之，数据模型就是

对现实世界抽象化的数据展示。数据模型在满足抽象的同时，越简单越好。

一种数据模型往往是为一种需求服务的，可能换个使用场景，就没有那么好的效果。拿业务数据表来说，一家创业公司为了满足正常的业务流程，往往会设计一套支撑正常业务流程的数据模型，这其中包括用户表、订单表、商品详情表等。随着公司发展，一两年后，就会有上百张表，这些表之间有比较复杂的依赖关系，如图 3-12 所示。

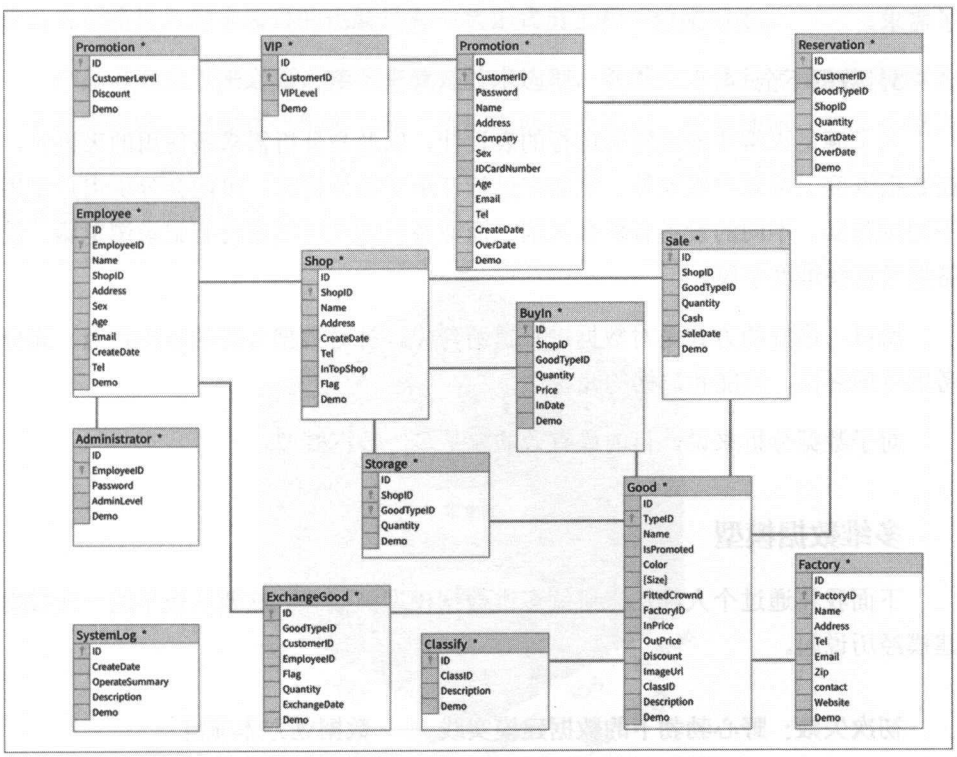


图 3-12　业务数据表

这套数据模型支撑正常的业务逻辑没有问题，但如果把它开放给业务人员进行数据分析，就会产生各种问题，具体来说有三点。

1. 数据报表晦涩难懂

业务数据表的设计，是为了业务系统的正常运转，是给机器而不是给人看的。虽然我们可以基于业务数据表进行一些统计分析，但可能只有开发人员和运维人员知道表的结构含义。对于业务人员来说，理解起来非常困难，并且业务数据表

可能会随着业务的发展进行一些增删字段、拆表之类的变更，使之非常难以驾驭。

2．性能问题

正常的业务流程，对于数据库中的数据更新具备小批量、高并发的特点，比如更新订单 008 的状态为已支付。而我们做数据分析时，查询的并发量并不会特别大，但分析的数据规模往往很大，比如最近半年有超过三次购买的男性用户的平均客单价，这类需求计算量很大，在业务数据库上进行查询的性能往往满足不了需求。

3．数据不全

为了满足正常业务运转而进行的表设计，以及与分析需求所使用的表设计，在数据维度上可能产生差异。比如在进行订单交易分析时，可能会分析用户使用不同浏览器、不同的设备有什么区别，而业务数据库可能就没有记录浏览器、设备型号这些维度字段。

所以，最好的方式是对数据进行重新建模，针对数据分析的应用特点，充分考虑可理解性、性能和数据的完备性。

对于数据分析来说，目前最有效的就是多维数据模型。

多维数据模型

下面我将通过个人经历来讲解多维数据模型的概念。这要从较早的一次数据建模经历说起。

初次失败：野心勃勃下的数据建模实践——数据立方体项目

2010 年年初，百度地图团队的一位 PM 找到我，他给我演示了一份 PPT——某公司的一份统计分析系统的对外交流材料。他告诉我，厂长（百度 CEO 李彦宏）看到这份材料，觉得做得挺好，并希望我们也做一套。

我发现，这个统计分析系统就是某个互联网产品的流量、用户量通过几个页面的展示，针对地域、渠道等几个维度进行分析。我当时想，这在我们的日志统计平台上很容易通过几个任务实现，但日志统计平台是以统计任务来管理的，虽然功能强大，但是不利于展示。对一个业务线来说，就是一组报表，并没有层级管理。相比之下，PPT 中演示的系统在界面组织上就好很多。于是我对这位 PM 说，

这套系统太简单了，既然我们要做，就要比他们做得好。我考虑了一下，然后给出一套方案。

就这样，我和团队的几个兄弟开始考虑如何做一套更好的方案。经过调研，我们发现数据仓库教材里介绍的数据立方体的模型，更适合做这件事。于是拿着这套方案和 PM 沟通，PM 听了介绍之后也赞叹不已。当时我相信自己要做的事情非常独特，超越之前的任何方案，却压根没有考虑人力是否能够支持，实际上最后真正能投入到本项目的也就只有一名正式员工和一位实习生。

产品方案定下来之后，接下来就是技术选型。数据立方体是多维数据模型的一个通俗叫法，主要由“维度”和“指标”两部分组成，比如地域、操作系统属于“维度”，销售额、注册用户数、成单量是“指标”。我们可以通过维度组合，查看该组合下的指标情况，如图 3-13 所示。

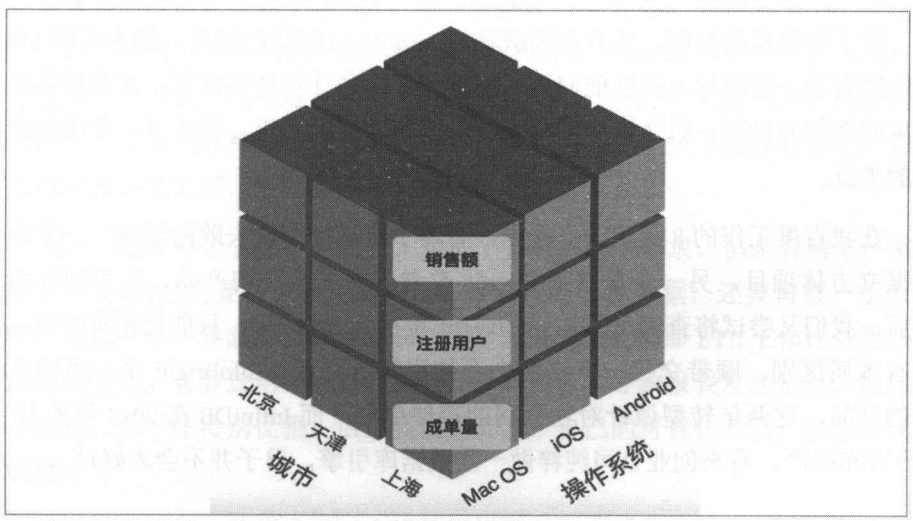


图 3-13 数据立方体的样例

这个模型非常清晰和简单，我们对百度进行流量分析，可以拆开多个维度，如时间、地域、渠道、操作系统、浏览器版本、频道、行为类型等。通过此数据立方体，我们可以查看北京用户使用 iOS 的销售额。

然而，该模型的难点在于数据规模。在单位时间内，所产生的数据条数就是所有维度的乘积，如果有 10 个维度，每个维度有 10 个项目，就会产生 10×10 条记录。如果每条记录按 1KB 大小计算，就有 10TB 数据量。显然，一台机器的性能是撑不住的。

我们继续寻找适合在我们数据规模上进行查询的存储系统，发现 InfoBright 这一存储引擎最合适，它采用列式存储，在针对多维数据分析这种模型上，性能很好。但因为是单机的，支持的数据规模有限，于是，我们对某些维度的元素进行了聚合，以降低数据量，最后降到半年的累计数据预计几百 GB。

就这样，我们在半个正式员工、两个实习生的人员配置下，野心勃勃地开启了整个项目。我们还把部门的高级总监邀请到开发群里——因为部门总监最需要针对流量数据的多维分析。

但是两个月后，结果并不尽如人意。

产品是做出来了，但多个维度的组合查询性能差得一塌糊涂，我在界面上进行某个查询，半个小时后还看不到结果，根本没法使用，整个产品只能算是个半吊子的 Demo，连部门总监也退出了群。

除了存储层的问题，还有查询解释层 Mondrian 的性能问题、报表引擎 JPivot 的性能问题、数据导入的性能问题、预处理数据的计算性能问题，以及数据字段变更的维护问题等。总之这个项目是在一个不合适的时机，提出了一个比较理想化的主意。

在我百度工作的 8 年职业生涯中，有两个我认为彻底失败的项目，一个就是数据立方体项目，另一个是基于 Impala 改进的交互式查询产品。认识到性能问题后，我们又尝试将查询引擎从 InfoBright 替换到 InfiniDB，只能说稍有改进，但没有本质区别。顺带交代一下这两个存储引擎的命运：InfoBright 是一家波兰公司的产品，这两年转型做针对物联网的存储引擎；而 InfiniDB 在 2014 年的 10 月 1 日宣布破产。看来创业公司纯粹做一款数据库引擎，日子并不会太好过。

渐入佳境：操刀 UDW，重新认知数据建模

这次项目失败后，我对数据立方体这种理论化的模型产生了怀疑，觉得在现实场景下实施有困难。又过了一年，百度成立了基础架构部数据团队，并从 Google 聘请了一位总监吕厚昌（Alex Lu），他可谓领域内的资深专家，就职百度之前，在 Yahoo! 工作 7 年，在 Google 工作 5 年，Google 的 Tenzing 引擎就是他的团队做出来的。

他来了之后，打开了我的思路，相比之下，我之前对数据架构的理解真的太狭隘了。他先是给我们提出了数据分层的金字塔模型，决定构建 UDW（User Data Warehouse），能够将用户在百度所有产品线的行为统一到一起去。有了这

个地基，剩下的数据使用问题，就变得容易了。

Alex Lu 给我讲解了在 UDW 基础之上，将用户数据按照时间细粒度汇聚，根据不同维度组合查询，所有的报表需求都产生在这个基础之上。相比之下，我们之前的报表数据，都是直接从原始数据计算生成统计结果，计算效率很低，而且中间数据没有得到重复使用。相比数据立方体项目，常规报表数据例行跑出，而不实时交互，这对查询性能要求没那么高。在 UDW 的基础之上，数据立方体的思路让我意识到竟然能很好地解决计算资源浪费的问题，十分惊叹。

对于交互式查询的需求，问题是一样存在的。我们数据团队是由两个团队合并创建的，一个是我所带领的数据平台团队，一个是内部叫 Doris 的分布式查询团队。Doris 主要解决海量数据下，使用 MPP 架构，满足毫秒级的查询问题（对外的百度统计以前就使用了这一系统）。把它改造一下，能够对接报表引擎即可。这个最重要的改造就是要支持 SQL。这一思路在一位 Google 架构师 James Peng 的加入后得以传递。Doris 团队的人员花了两周时间，直接将 Doris 作为 MySQL 的存储引擎，这样就可以通过 MySQL 直接访问 Doris，支持了 SQL 语法。Infobright 也是这么一个实现思路，于是查询性能的问题也解决了。所有的核心报表，都通过数据立方体来实现，展现部分用了 Oracle BIEE。

Oracle BIEE 基于多维数据模型，实现了报表的基本需求。但是有两个严重的问题，一是 BIEE 配置报表非常麻烦，即使是规整好的数据，还要再建一层数据模型，多此一举，界面操作非常复杂；二是数据的预处理即 ETL 工作比较复杂，数据源的变更，会导致结果出错，ETL 计算周期长，导致报表发送延迟。虽然能基本满足，但不是特别优雅。后来我们又开发了自己的可视化系统，解决报表展示问题。

通过以上讲解，希望读者对多维数据模型有更全面的认识。接下来，我们将介绍多维数据模型和用户行为分析中的行为事件相结合所发挥的威力。

多维事件模型

在介绍多维事件模型（Event 模型）之前，我们先来认识一下访问量模型。

访问量模型

在传统的 Web 时代，我们通常使用 PV 来衡量和分析一个产品的好坏。而在

移动互联网及O2O电商时代，PV已经远远不能满足产品和运营人员的分析需求。

每个产品都有独一无二的核心指标，用来衡量产品是否成功，这个指标也许是发帖数量、视频播放数量、订单量或者其他可以体现产品核心价值的指标，这些都是一个简单的PV无法衡量的。

除此之外，PV模型也无法满足一些更加细节和精细化的分析。例如，我们想分析哪类产品销量最好，访问网站的用户年龄和性别构成，其他渠道的用户转化率、留存和重复购买率如何，新老用户的客单价、流水、补贴比例分别是多少等。这些都是以PV为核心的传统统计分析没办法解答的问题。

多维事件模型

多维事件模型分成Event实体和User实体。

1．Event实体

简单来说，Event描述了一个用户在某个时间点、某个地方以某种方式完成某个具体事情。从这可以看出，一个完整的Event，包含如下的几个关键因素。

• Who：即参与这个事件的用户是谁。在我们的数据接口中，使用distinct_id来设置用户的唯一ID：对于未登录用户，这个ID可以是Cookie、设备ID等匿名ID；对于登录用户，则建议使用后台分配的实际用户ID。同时，神策分析也提供了track_signup这个接口，在用户注册的时候调用，用来将同一个用户注册之前的匿名ID和注册之后的实际ID贯通起来进行分析。

• When：即这个事件发生的实际时间。在数据接口中，使用time字段来记录精确到毫秒的事件发生时间。如果调用者不主动设置，则各个SDK会自动获取当前时间作为time字段的取值。

• Where：即事件发生的地点。使用者可以设置properties中的$IP属性，这样系统会自动根据IP来解析相应的省份和城市，当然，使用者也可以根据应用的GPS定位结果，或者其他方式来获取地理位置信息，然后手动设置$city和$province。除了$city和$province这两个预置字段以外，也可以自己设置一些其他地域相关的字段。例如，从事社区O2O的产品经理，可能需要关心每个小区的情况，则可以添加自定义字段“HousingEstate”；从事跨国业务的产品经理，需要关心不同国家的情况，则可以添加自定义字段“Country”。

• How：即用户从事这个事件的方式。这个概念比较广，包括用户使用的设备、使用的浏览器、使用的 APP 版本、操作系统版本、进入的渠道、跳转过来时的 Referer 等。目前，神策分析预置了如下字段用来描述这类信息，使用者也可以根据自己的需要来增加相应的自定义字段，如表 3-1 所示。

表 3-1 预置字段对应描述信息

字 段	描 述
$ APP_version	应用版本
$ city	城市
$ manufacturer	设备制造商，字符串类型，如“Apple”
$ model	设备型号，字符串类型，如“iPhone6”
$ os	操作系统，字符串类型，如“iOS”
$ os_version	操作系统版本，字符串类型，如“8.1.1”
$ screen_height	屏幕高度，数字类型，如 1920
$ screen_width	屏幕宽度，数字类型，如 1080
$ wifi	是否 WIFI，BOOL 类型，如 true

• What：描述用户所做的这个事件的具体内容。在数据接口中，首先使用“Event”这个事件名称来对用户所做的内容进行初步分类。Event 的划分和设计也有一定的指导原则，我们会在后文详细描述。除了“Event”这个至关重要的字段以外，我们并没有设置太多预置字段，而是请使用者根据每个产品以及每个事件的实际情况和分析的需求，来进行具体的设置，如表 3-2 所示。

表 3-2 设置事件字段

事 件	记录字段
购买	商品名称、商品类型、购买数量、购买金额、付款方式等
搜索	搜索关键词、搜索类型等
点击	点击 URL、点击 title、点击位置等
用户注册	注册渠道、注册邀请码等
用户投诉	投诉内容、投诉对象、投诉渠道、投诉方式等
申请退货	退货金额、退货原因、退货方式等

2. User 实体

每个 User 实体对应一个真实的用户，用 distinct_id 进行标识，描述用户的长期属性（也即 Profile），并且通过 distinct_id 与这个用户所从事的行为，也即 Event 进行关联。

一般记录 User Profile[1] 的场所，是用户进行注册、完善个人资料、修改个人资料等几种有限的场合，与 Event 类似，建议在后端记录和收集 User Profile。

收集哪些字段作为 User Profile，也完全取决于产品形态及分析需求。简单来说，就是在能够拿到的那些用户属性中，哪些对于分析有帮助，则作为 Profile 进行收集。图 3-15 是针对某知名直播平台业务需求进行的数据模型设计。

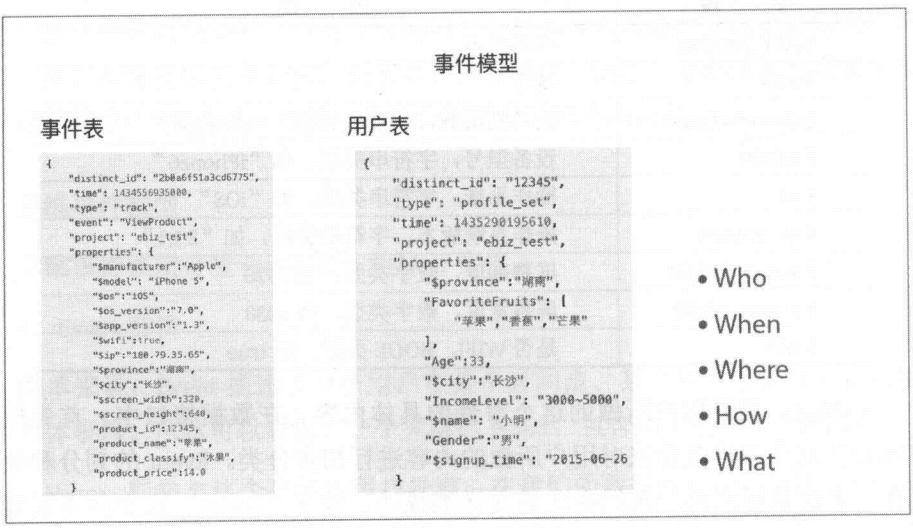


图 3-15　某知名直播平台案例之事件模型

多维事件模型的探索经历

多维事件模型的探索经历，要从用户行为分析想法的诞生说起。

2008 年，也就是我在百度知道工作一年的时间中，我们通过问题推荐的项目给产品带来了 7.5% 的回答量提升，之后发现很难找到发力点。百度知道从 2005 年推出，到那时已经过了三年时间，进入了成熟期，很难再通过某个功能改进获得产品数据的大幅提升。我想，能不能借助用户行为分析来了解百度知道的用户行为特征，也许我们都不了解用户使用产品的方式，但可以通过用户行为分析发现不同用户的访问特征，进而优化产品。

那时我对用户行为分析的理解很初级，直觉是应该做一个产品——能够研究用户的访问行为序列。我先是拿了一些 Web Server 的日志，然后人工去抽查，但浏览起来太不方便。我想，如果有个工具，在输入一个用户账号后就能看到他的

[1] User Profile，用户画像，第 5 章将对此进行深入讲解。

详细行为，这会方便很多。

不久，我去负责日志统计平台，这个工作就暂停了。到了2009年中期，日志统计平台已经顺利推出，并且在公司内得到很好的推广，于是我将一些研发工作都移交给团队新人，又抽出手来继续研究用户行为分析。我的思路得到了领导的认可，新产品部门把几个业务线的人联合起来，借调了三四个工程师，作为一个跨团队的项目开展。我的思路是要把用户在百度知道、贴吧、图片、MP3等这些产品的用户行为全部规范化，导入到一个系统中，通过一个界面能够查询到任意指定用户的详细行为序列。我当初把这个平台定义为“用户行为分析平台”，可部门的架构师听了之后，觉得这个名字太大了，因为这只是一个用户行为查询平台。

用户行为查询平台

我把用户的任意一个行为操作，都分配一个行为ID，叫ActionID，它是一个数字串，由两部分组成，前三位是产品号，后三位是行为编号，比如百度知道提问行为是101001的形式。每个行为都包含一系列的属性，包括提问行为的用户ID、浏览器类型、IP、问题标题等。不同的行为，有通用的属性，也有特有的属性。有些属性需要和产品线建立查询接口来获取，访问日志中是不包含的。现在想想这个查询接口的设计很不合理，因为它让业务线的耦合性太强。

整个平台主要有两个关键点，一是数据的生成，二是数据的存储和查询。我们直接在强大的日志统计平台上开发几个MapReduce任务，将Web Server及其他后端模块的日志，处理为以ActionID为核心的行为数据，以完成数据的生成。对于数据的存储和查询，我们采用了Hypertable，当时百度系统部有三个工程师在研究Hypertable，并在社区贡献代码。Hypertable是Google BigTable论文的一个开源实现，由C++开发而成，现在用的人已经很少了。我们把处理好的用户行为数据导入Hypertable，并提供一个查询的界面，输入一个IP、UserID之类的内容就可以查看用户的行为序列。

平台推出后，我发现没什么人用。这只是一个内部的工具，主要面向产品经理，他们只会偶尔尝尝鲜，并没有很强的应用场景。与此同时，我也发现网页搜索部有个类似的工具用来人工评估一些检索的用户满意度，网页搜索的产品经理用得挺多。

LogData 平台

2010 年 4 月，网页搜索部和搜索新产品部合并了，我们的日志统计团队和网页搜索部的用户行为分析团队合并。我们就把原网页搜索部开发的一些查询工具，全部统一到一起，叫 LogData 平台。许多查询工具都挺类似，一个 ID 对应许多条记录，所以很容易统一，这样就避免每产生一个需求，都单独开发和维护一套新工具的问题。

在这之前，负责 Hypertable 的系统部同事已经离职，加上开源社区都转向了 HBase，Hypertable 本身一直没有到达稳定的状态，而且三天两头出现故障，需要重新导入数据。于是我们也抛弃了 Hypertable 的使用，开始转向 HBase。刚开始用 HBase 时也经常出问题，经过几个月的完善，算是比较稳定了。谁知维护 HBase 的系统部团队又一次出现倒戈，开始推进内部研发的一个叫 DDBS 的分布式 MySQL。DDBS 本来是为内部的广告系统而研发的，广告部门的同事没有采纳，反而我们成了第一大用户。DDBS 的存储效率不如 HBase，有一个好处就是可以用 SQL 进行查询海量数据。

用户行为数据

前面提到的用户行为查询平台的使用情况并没那么乐观，但我们生成的用户行为数据反而有了更多的应用场景。

一是用于用户行为序列挖掘的研究，当时我花了不少时间研究时间序列的论文，并进行了一些序列模式挖掘的尝试，通过把用户的 ActionID 序列抽取出来进行。二是我和另一同事两个人针对百度知道、百度百科、百度图片等业务线，通过用户行为数据进行了许多用户分群的挖掘尝试，比如把访问频道和新老用户组合起来分析，生成分析报告。同样，这些分析报告发现不能直接给产品线带来价值，许多产品经理看了后，也并不能指导他的实际行动。现在看来，若利用神策分析的多维事件分析功能，这些组合分析都非常简单，当时我们可花费了不少工作和心血。

还有一个应用点是我们的用户行为数据成为个性化推荐的基础，2010 年，百度成立了专门的推荐团队，专门研究推荐引擎，我们的数据成了重要数据源，在我离职前还在应用。

2011 年下半年，我们开展用户数据仓库项目，同样是做用户行为事件的分析，在行为数据的整理上，最大的区别是 ActionID 改叫 EventID，基本理念是一样的。再之后我通过 Event 模型将全百度的数据进行统一，叫 User Data Warehouse。在

百度的几年时间里把 Event 模型在百度公司发挥到比较理想的状态，我也深刻理解了这一模型在用户行为分析上的强大之处。

数据分析方法

基于多维事件模型，会形成一些常见的数据分析方法。在用户行为领域，对这些数据分析方法的科学应用进行理论推导，能够相对完整地揭示用户行为的内在规律。基于此帮助企业实现多维交叉分析，让企业建立快速反应、适应变化的敏捷商业智能决策。

接下来我们将为大家逐一介绍常见的数据分析方法。值得强调的是，每一种数据分析方法都针对不同维度的数据研究。各分析模型存在相互依赖的关系，精益数据分析是数据分析方法交叉应用的结果。

行为事件分析

接下来，我们分别从行为事件分析的定义、特点与价值，以及应用场景几个方面进行介绍。

行为事件分析的定义

行为事件分析法用来研究某行为事件的发生对企业组织价值的影响以及影响程度。企业借此来追踪或记录用户行为或业务过程，如用户注册、浏览产品详情页、成功投资、提现等，通过研究与事件发生关联的所有因素来挖掘用户行为事件背后的原因、交互影响等。

在日常工作中，运营、市场、产品、数据分析师根据实际工作情况而关注不同的事件指标。如最近三个月来自哪个渠道的用户注册量最高？变化趋势如何？各时段的人均充值金额是分别多少？上周来自北京发生过购买行为的独立用户数，按照年龄段的分布情况如何？每天的独立 Session 数是多少？诸如此类的指标在查看的过程中，行为事件分析起到重要作用。

行为事件分析涉及事件、维度和指标三个概念。在分析过程中，一般期望数据是实时采集并能够实时分析的，而事件、维度和指标是可以灵活自定义的。行为事件分析是上节内容讲到的 Event 实体的可视化展现，其中还将 User 实体的属

性通过 User ID 贯穿到 Event 实体中，这样在分析时可以把用户属性作为分组或筛选的条件。

行为事件分析模型的特点与价值

行为事件分析法具有强大的筛选、分组和聚合能力，逻辑清晰且使用简单，已被广泛应用。行为事件分析法一般经过事件定义与选择、多维度下钻分析、解释与结论等环节。

• 事件定义与选择。事件描述的是一个用户在某个时间点、某个地方、以某种方式完成了某个具体的事情。Who、When、Where、What、How 是定义一个事件的关键因素。其中：

Who 是参与事件的主体，对于未登录用户，可以是 Cookie、设备 ID 等匿名 ID，对于登录用户，可以使用后台配置的实际用户 ID。

When 是事件发生的实际时间，应该记录精确到毫秒的事件发生时间。

Where 即事件发生的地点，可以通过 IP 来解析用户所在省市，也可以根据 GPS 定位方式获取地理位置信息。

How 即用户从事该事件的方式，包括用户使用的设备、浏览器、APP 版本、渠道来源等。

What 描述用户所做该事件的所有具体内容。比如对于“购买”类型的事件，需要记录的字段有商品名称、商品类型、购买数量、购买金额、付款方式等。

• 多维度下钻分析。高效的行为事件分析要支持任意下钻分析和精细化条件筛选。当行为事件分析合理配置追踪事件和属性，可以激发出事件分析的强大潜能，为企业回答关于变化趋势、维度对比等各种细分问题。同时，还可以通过添加筛选条件，精细化查看符合某些具体条件的事件数据。

• 解释与结论。此环节要对分析结果进行合理的理论解释，判断数据分析结果是否与预期相符，如判断产品的细节优化是否提升了触发用户数。如果相悖，则应该针对不足的部分进行再分析与实证。

行为事件分析的应用场景

该场景为互联网金融行业常见的行为事件分析应用场景。

某互联网金融客户运营人员发现，4月10日来自新浪渠道的PV数异常标高，因此快速排查原因：是异常流量还是虚假流量？

企业可以先定义事件，通过“筛选条件”限定广告系列来源为“新浪”。再从其他多个维度进行细分下钻，比如“地理位置”“时间”“广告系列媒介”“操作系统”“浏览器”等。当进行细分筛查时，虚假流量无处遁形。图3-16为来源于“新浪”的各城市浏览页面的总次数。

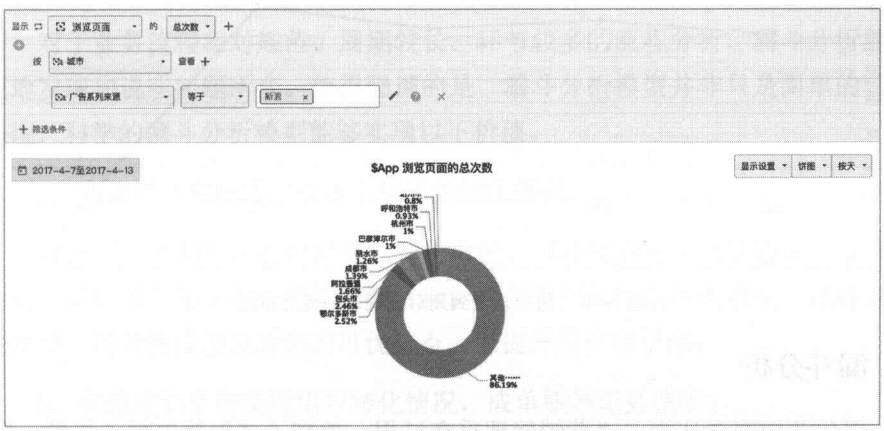


图3-16　来源新浪的各城市浏览页面的总次数

在剔除虚假流量后，运营人员可进行其他用户行为分析。通过“投资成功”事件，查看各个时段的投资金额。若想知道每个产品类型的投资金额，此时再按照“产品类型”进行分组查看即可。图3-17显示不同产品投资成功的支付金额的总和。

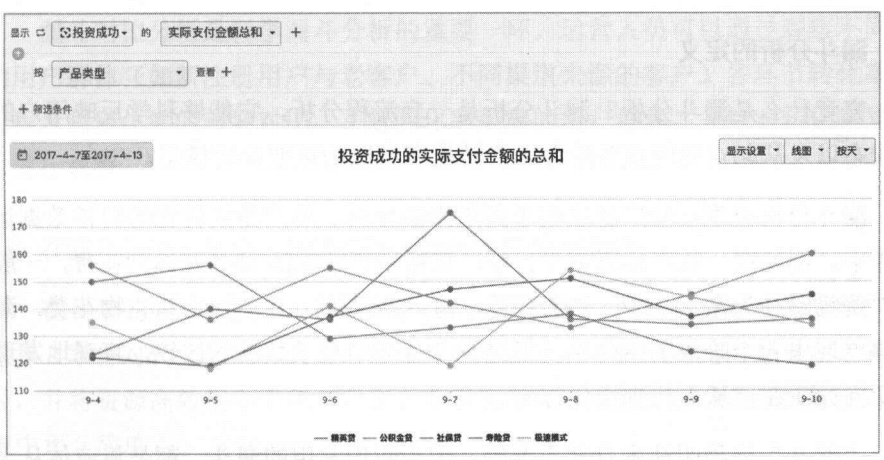


图3-17　不同产品投资成功的支付金额的总和

当用户投资到期后，后续行为可能是提现或继续投资，运营人员可以实时关注“提现率”的变化趋势，如图 3-18 所示。

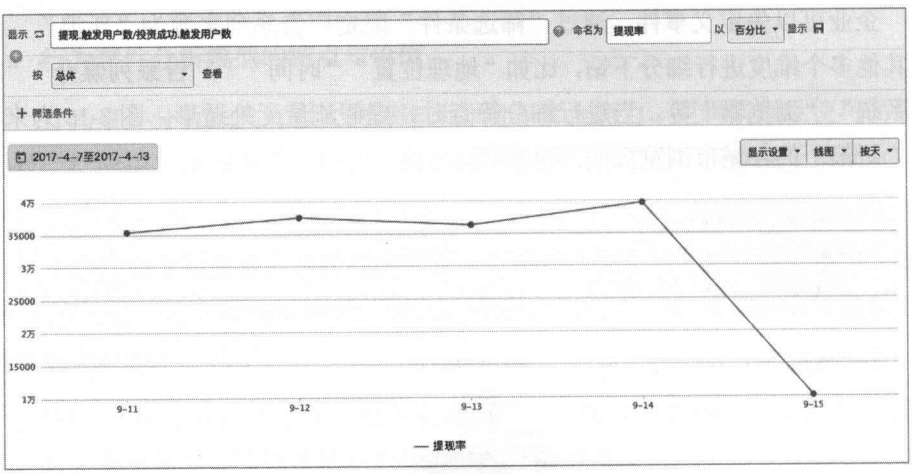


图 3-18　用户投资到期后提现率的变化情况

漏斗分析

现代营销观念认为：“营销管理重在过程，控制了过程就控制了结果。”漏斗分析模型是企业实现精细化运营的重要分析模型，其精细化程度影响着营销管理的成败。粗陋的漏斗分析模型因为“过程管理不透明”造成结果失控。因此，我们经常能够听到一些产品经理的抱怨：从启动 APP 到“支付成功”，用户转化率为何仅仅 0.8%？

漏斗分析的定义

究竟什么是漏斗分析？漏斗分析是一套流程分析，它能够科学反映用户的行为状态以及从起点到终点各阶段用户转化率情况的重要分析模型。

漏斗分析模型已经广泛应用于渠道来源分析、用户激活转化等日常数据运营工作中。例如在一款产品服务平台中，直播用户从激活 APP 开始到花费，一般的用户购物路径为激活 APP、注册账号、进入直播间、互动行为和礼物花费，漏斗能够展现出各个阶段的转化率，通过漏斗各环节相关数据的比较，直观地发现和说明问题所在，从而找到优化方向。

“漏斗”这种叫法本身并不准确，因为我们常用的漏斗，都是将液体从大开

口导入，从小开口漏出的，最终所有液体流出。但我们所说的漏斗分析，每个前序步骤只会有部分用户进入到下一步骤中，所以我觉得叫“漏筛”更准确一些。这里需要注意的是，我们跟踪整个漏斗的转化过程，是以用户为单位将步骤串联起来，并不是只把每个步骤的发生次数做一个简单的计数，进入到后续步骤中的用户，一定是完成了所有前序步骤。

漏斗分析模型的特点与价值

对于业务流程相对规范、周期较长、环节较多的流程分析，漏斗分析能够直观地发现和说明问题所在。值得强调的是，漏斗分析模型并非只是简单的转化率呈现，科学的漏斗分析模型能够实现以下价值。

1．企业可以监控用户在各个层级的转化情况。

聚焦用户选购全流程中最有效转化路径，同时找到可优化的短板，提升用户体验。降低流失是运营人员的重要目标，通过不同层级的转化情况，迅速定位流失环节，针对性持续分析找到可优化点，以提升用户留存率。

2．多维度切分与呈现用户转化情况，成单瓶颈无处遁形。

科学的漏斗分析能够展现转化率趋势的曲线，帮助企业精细地捕捉用户行为变化。提升了转化分析的精度和效率，对选购流程的异常定位和策略调整效果验证有科学指导意义。

3．不同属性的用户群体漏斗比较，从差异角度窥视优化思路。

漏斗对比分析是科学漏斗分析的重要一环。运营人员可以通过观察不同属性的用户群体（如新注册用户与老客户、不同渠道来源的客户）各环节转化率，各流程步骤转化率的差异对比，了解转化率最高的用户群体，并针对转化率异常环节进行调整。

在漏斗分析方法中，科学归因、属性关联的重要性

在科学的漏斗分析中，需要科学归因设置。每一次转化节点应根据事件功劳差异（事件对转化的功劳大小）而科学设置。企业一直致力定义最佳用户购买路径，并将资源高效集中于此。而在企业真实的漏斗分析中，业务流程转化并非理想中那么简单。

以市场营销为例，市场活动、线上运营、邮件营销都可能触发用户购买。A欲选购一款化妆品，通过市场活动了解M产品，又在百度贴吧了解更多信息，但是始终没有下定决心购买。后来收到M公司的营销邮件，A被打折信息及详实的客户评价所吸引，直接邮件内跳转至网站购买了该商品。

那么，在漏斗设置时，转化归因应该“归”哪一个渠道呢？在这个案例中，运营人员愿意以实际转化事件的属性为准。邮件营销的渠道在用户购买决策的全流程中对用户影响的“功劳”最大、权重较大，直接促进用户转化。在科学的漏斗分析模型中，用户群体筛选和分组时，以实际转化事件——邮件营销来源的用户群体的属性为准，大大增大了漏斗分析的科学性。

此外，在进行漏斗分析时，尤其电商行业的数据分析场景中，运营人员在定义“转化”时，会要求漏斗转化的前后步骤有相同的属性值。比如同一ID（包括品类ID、商品ID）才能作为转化条件——浏览iPhone 6，购买同一款iPhone 6才能被定义为一次转化。因此，“属性关联”的设置功能是科学漏斗分析不可或缺的内容。

漏斗分析模型的应用场景

某电商企业客户根据客户的消费能力，将客户划分为普通会员、黄金会员和钻石会员。为加强对用户的转化引导，企业欲针对不同用户群体采用不同的运营方式。

如图3-19所示，通过对比可明显看出，普通会员从“提交订单”到“支付订单”的转化率明显低于钻石会员。为找到“支付订单”阶段转化率变低的原因，企业运营人员应深度分析普通会员转化率情况，如对比不同付费渠道（PC端、移动端等）的转化情况，找到优化的短板，比如尝试支付订单流程的新手引导，帮助新手顺利完成购买。

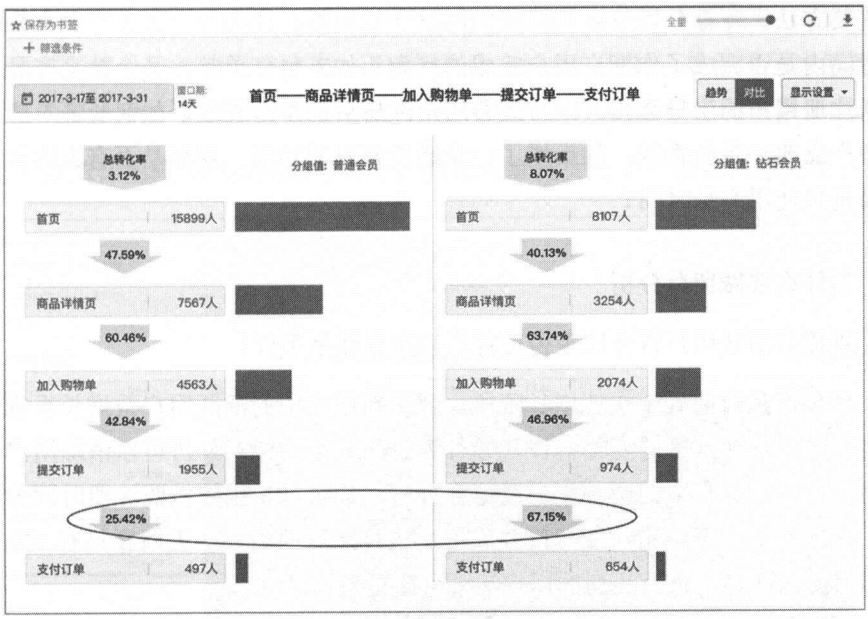


图 3-19　普通会员与钻石会员的漏斗转化情况对比

此外，我们还可以选择某一转化步骤，从而拿到这一步骤流失的用户列表，对其进行运营，以提升转化。

留存分析

据某第三方平台近期调研结果显示，在金融创业领域，2013 年某家互联网金融创业公司的投资获客成本区间为 300 ～ 500 元，而 2016 年则涨为 1000 ～ 3000 元；在电商领域，新用户的获取成本，是维护一个老用户的 3 ～ 10 倍……

如今，高居不下的获客成本让互联网、移动互联网创业者们遭遇新的“天花板”，甚至陷入新客获取难的窘境。花费极高成本所获取的客户，可能仅打开一次 APP 或完成一次交易后就白白流失。随着市场饱和度上升，绝大多数企业亟待解决如何增加客户黏性，延长每一个客户的生命周期价值的问题。因此留存分析这一分析模型备受青睐。

留存分析的定义

留存分析是一种用来分析用户参与情况和活跃程度的分析模型，考察进行初始行为的用户中，有多少人会进行后续行为。这是用来衡量产品对用户价值高低

的重要方法。留存分析可以帮助我们回答一些问题，比如一个新客户在未来的一段时间内是否完成了你期许用户完成的行为？如支付订单等；某个社交产品改进了新注册用户的引导流程，期待改善用户注册后的参与程度，如何验证？想判断某项产品改动是否奏效，如新增了一个邀请好友的功能，观察是否有人因新增功能而延长使用产品时间？

为什么要做留存分析

直接看活跃用户百分比是否可行？答案显然是不行！

如果产品目前处于快速增长阶段，那么新用户中的活跃用户数增长很有可能掩盖了老用户活跃度的变化。按初始行为时间分组的留存分析可以消除用户增长对用户参与数据带来的影响。通过留存分析，你可以将用户按照注册时间分段查看，得出类似如下结论：三月份改版前，该月注册的用户 7 天留存只有 15%，但是四月份改版后，该月注册的用户 7 天留存提高到了 20%。

留存分析模型特点与价值

科学的留存分析模型具有灵活条件配置的特点 —— 根据具体需求筛选初始行为或后续行为的细分维度，针对用户属性筛选合适的分析对象。那么，留存分析有哪些价值呢？

1. 留存率是判断产品价值最重要的标准，揭示了产品保留用户的能力。

留存率反映的是一种转化率，即由初期不稳定的用户转化为活跃用户、稳定用户、忠诚用户的过程。随着统计数字的变化，运营人员可看到不同时期用户的变化情况，从而判断产品对客户的吸引力。

2. 宏观把握用户生命周期长度及定位产品可改善之处。

通过留存分析，我们可以查看新功能上线之后，对不同群体的留存是否带来不同效果？判断产品新功能或某活动是否提高了用户的留存率？结合版本更新、市场推广等诸多因素，去除使用频率低的功能，实现快速迭代验证，制定相应的策略。

留存分析模型的应用场景

场景 1：游戏行业提升活跃、留存 —— 如何精准找到玩家“流失点”

游戏的生命周期的时长差异和玩家的游戏黏度，直接体现了游戏的竞争能力

和赢利能力。玩家对游戏的直观感受、游戏难度曲线、游戏节奏的松弛、游戏福利等因素都能够导致游戏玩家流失。正确找到玩家流失的原因，是促进玩家活跃、挽留玩家的第一步。下面是游戏《迷城物语》在测试期间的相关应用情景。

图 3-20 统计出流失玩家的等级分布，判断玩家流失与关卡设置的相关性。

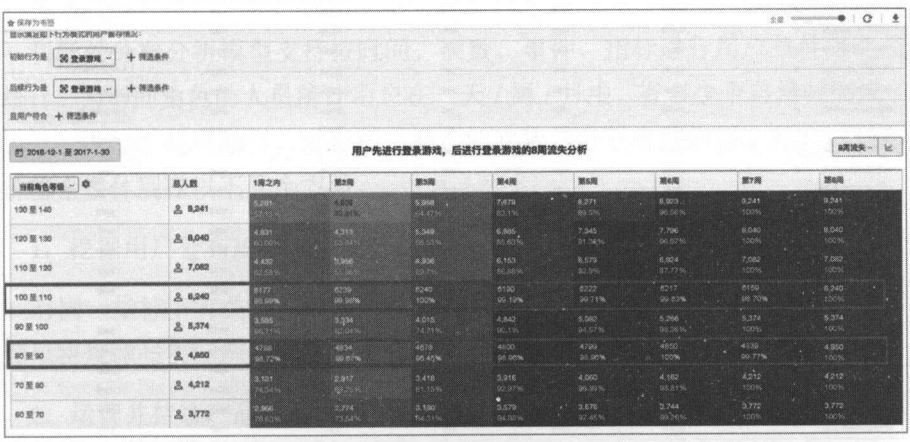


图 3-20　玩家在首次登录游戏之后的 8 周流失情况分析

其中 100 ～ 110、80 ～ 90 级是玩家流失较多的关卡。为精准定位导致玩家流失的关键因素，需要每个环节、具体场景进行深入追踪与分析。

场景 2：如何了解新用户的留存

运营人员想从总体上看用户留存的情况是否越来越好了。可根据新用户启动 APP 的时间按日或按月进行分组，观察该群体用户发生投资的 7 日留存、14 日留存或 30 日留存（可自由选择），点击“曲线标识”按钮，也可以看到每天留存率的变化趋势。如图 3-21 所示。

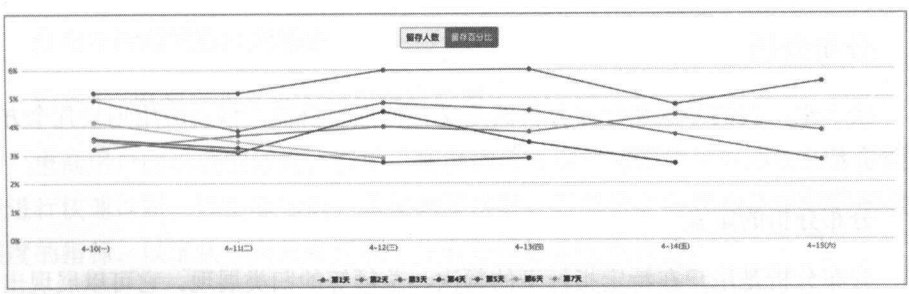


图 3-21　新用户群体 7 天留存趋势变化

7 日或者 30 日仍留下来做投资的用户，显然是一批忠诚度非常高的用户。以 4 月 10 日这天为例，一共有 1931 个新用户，在第 7 天有 68 人留了下来，用户列表界面如图 3-22 所示。

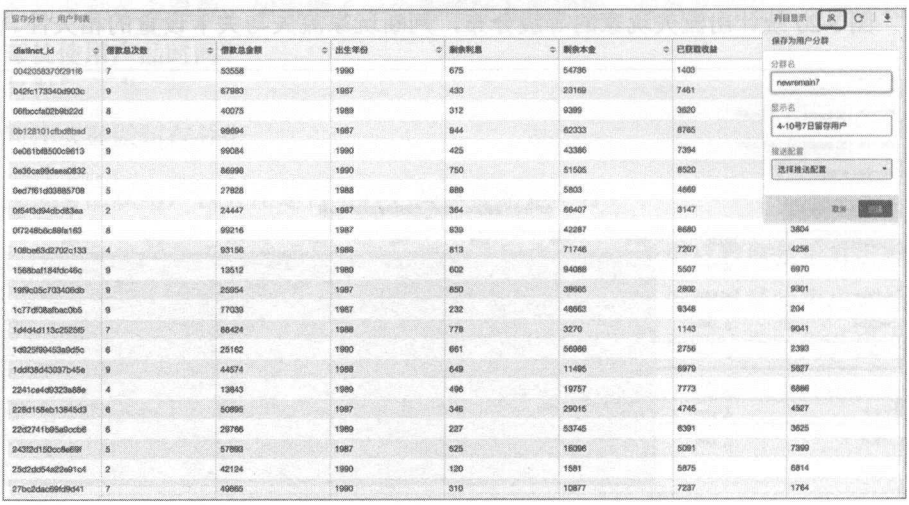


图 3-22　第 7 天用户留存中 68 人基本信息明细

我们能够看到留存下来用户的一些详细的基础信息，比如借款次数、借款金额、年龄等，通过总借款次数及借款金额，进行用户质量评估，通过年龄可以分析金融平台吸引的群体用户的年龄分布。另外，值得强调的是，支持人群明细查看是数据分析方法中不可或缺的功能。

若想深度挖掘高留存用户有哪些共性特征和他们的具体操作流程，以作为后序产品优化与改进的借鉴，则可使用用户分群功能，命名为“4 ～ 10 日 7 日留存用户”，然后通过用户路径等其他分析模型进一步深度分析。

分布分析

接下来，我们分别从分布分析的定义、特点与价值，以及应用场景几个方面进行介绍。

分布分析的定义

分布分析是用户在特定指标下的频次、总额等的归类展现。它可以展现出单用户对产品的依赖程度，分析客户在不同地区、不同时段所购买的不同类型的产

品数量、购买频次等，帮助运营人员了解当前的客户状态，以及客户的运转情况。如订单金额（100 以下区间、100 元 ～ 200 元区间、200 元以上区间等）、购买次数（5 次以下、5 ～ 10 次、10 以上）等用户的分布情况。

分布分析模型的特点与价值

科学的分布分析模型支持按时间、次数、事件、指标进行用户条件筛选及数据统计。为不同角色的人员统计用户在一天 / 周 / 月中，有多少个自然时间段（小时 / 天）进行了某项操作、进行某项操作的次数、进行事件指标。总之，分布分析价值主要体现在以下几个方面。

1. 挖掘用户分布规律，优化产品策略。

对同一指标下有关数据的统计与分析，帮助企业从中挖掘用户访问规律，企业可以将规律与实际产品策略相结合，进一步修正和重新制定产品策略。

2. 运营并持续产品生命力，增加客户回访率。

彻底改变之前依靠随机抽样的回访率调查方式，如电话回访等，分布分析从多角度分析辅助企业，判断单用户对产品的依赖程度，以及产品对用户的价值与黏性。

3. 快速识别核心用户群体，资源配置有的放矢。

核心用户群体是对企业价值贡献最大的用户群体，是企业最大的利润来源。不同用户群体对产品需求不一样，对用户群体进行差异性辨识，可以了解到用户群体对产品的依赖动力。分布分析通过不同维度筛选出核心用户群体，在此基础上，更好地配置优质资源，以最小成本实现企业利润最大化。

分布分析模型的应用场景

场景 1：电商行业常见的分布分析应用

电商用户的忠诚度如何、客单价情况如何等这些问题均可以通过分布分析功能进行快速诊断。以电商为例，重复购买次数、客单价分布等均是常用的衡量忠诚度的指标。以下从不同角度展现了分布分析的多维度查看。

图 3-23 可以看到用户每个月的购买频次基本稳定在 1 ～ 3 次之间，3 月有小的变动，其他几个月都比较稳定。

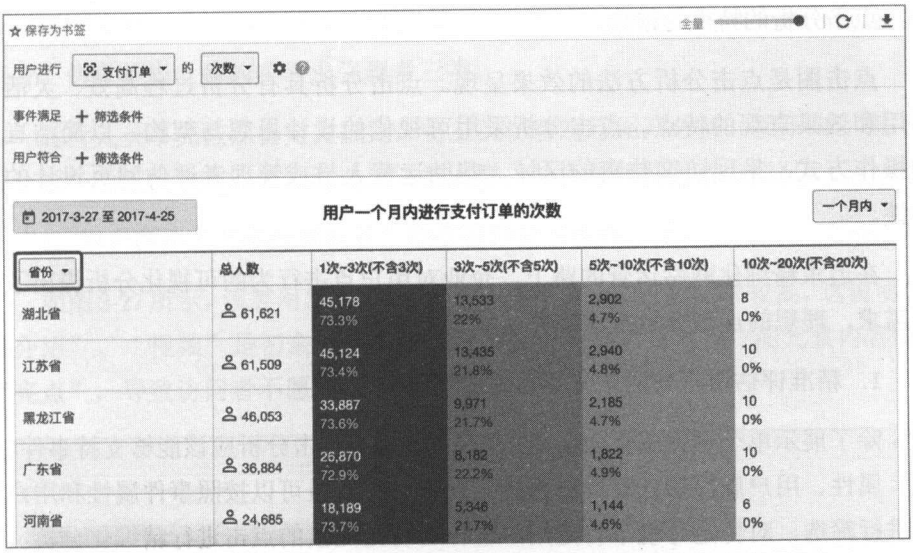
除了从用户行为日期去查看外，还可以对用户进一步细分，看看不同性别、不同渠道的用户的支付频次的差异，如图 3-24 所示，从性别来看，数据比较均匀。

图 3-23　用户在购买一个月内进行支付订单的次数

图 3-24　用户一个月内进行支付订单的次数

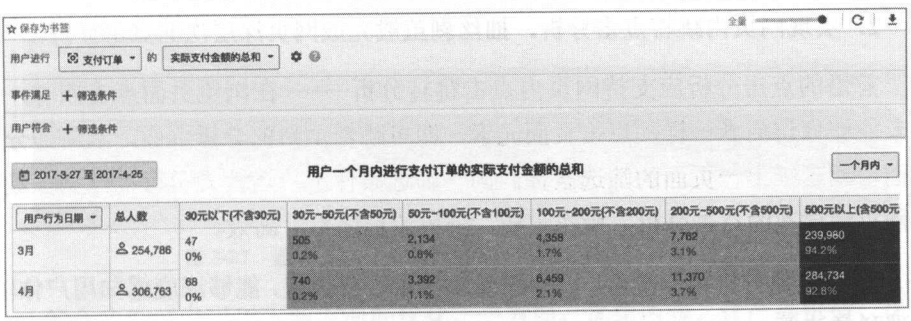
如图 3-25 所示，从省份的角度看，该商品并没有地域偏爱。

用户购买的客单价分布在哪个区间，也是运营人员比较关心的。图 3-26 显示出用户的客单价很高，90% 的用户客单件在 500 元以上。



省份	总人数	1次~3次(不含3次)	3次~5次(不含5次)	5次~10次(不含10次)	10次~20次(不含20次)
湖北省	61,621	45,178 73.3%	13,533 22%	2,902 4.7%	8 0%
江苏省	61,509	45,124 73.4%	13,435 21.8%	2,940 4.8%	10 0%
黑龙江省	46,053	33,887 73.6%	9,971 21.7%	2,185 4.7%	10 0%
广东省	36,884	26,870 72.9%	8,182 22.2%	1,822 4.9%	10 0%
河南省	24,685	18,189 73.7%	5,346 21.7%	1,144 4.6%	6 0%

图 3-25 按省份查看用户一个月内支付订单次数



用户行为日期	总人数	30元以下(不含30元)	30元~50元(不含50元)	50元~100元(不含100元)	100元~200元(不含200元)	200元~500元(不含500元)	500元以上(含500元
3月	254,786	47 0%	505 0.2%	2,134 0.8%	4,358 1.7%	7,762 3.1%	239,980 94.2%
4月	306,763	68 0%	740 0.2%	3,392 1.1%	6,459 2.1%	11,370 3.7%	284,734 92.8%

图 3-26 用户一个月内实际支付订单金额总和

点击分析

接下来，我们分别从点击分析的定义、特点与价值，以及应用场景几个方面进行介绍。

点击分析的定义

点击分析，即应用一种特殊高亮的颜色形式，显示页面或页面组（结构相同的页面，如商品详情页、官网博客等）区域中不同元素点击密度的图示。包括元素被点击的次数、占比、发生点击的用户列表、按钮的当前与历史内容等因素。

点击分析的特点与价值

点击图是点击分析方法的效果呈现。点击分析具有分析过程高效、灵活、易用和效果直观的特点。点击分析采用可视化的设计思想与架构，以简洁直观的操作方式，呈现访客热衷的区域，帮助运营人员或管理者评估网页设计的科学性。

在追求精细化网站运营的路上，企业对用户点击行为的可视化分析提出了更高需求，理想的点击分析方法能够实现以下价值。

1．精准评估用户与网站交互背后的深层关系。

除了展示单个页面或页面组的点击图，前沿的点击分析应该能够支持事件（元素）属性、用户属性的任意维度筛选下钻；运营人员可以按照事件属性和用户属性进行筛选，对特定环境下特定用户群体对特定元素的点击进行精细化分析；支持查看页面元素点击背后的用户列表，满足企业网站的精细化分析需求。

2．实现网页内跳转点击分析，抽丝剥茧般完成网页深层次的点击分析。

前沿的点击分析应支持网页内点击跳转分析 —— 在浏览页面点击图时，使用者能够像访问者一样，点击页面元素，即可跳转至新的分析页面，且新的分析页面自动延续上一页面的筛选条件。同一筛选条件下，运营人员可抽丝剥茧般完成网页深层次的点击分析，操作流畅，分析流程简易、高效。

3．与其他分析模型配合，以全面视角探索数据价值，能够深度感知用户体验，实现科学决策。

无法精细化地深入分析，会让网页设计与优化丧失了科学性。点击图呈现用户喜爱点击的模块或聚焦的内容，是数据价值最上层表现。当点击分析与其他分析模块配合，交叉使用，将数据和分析结果以多种形式可视化展现，运营人员即可深度感知用户体验。例如，改版后，如何评估新版本对用户体验的影响？一处修改，是否影响其他元素的点击？或通过 A / B 测试，反复验证优化效果选择最优方案等。

点击分析应用场景

场景 1：企业官网改版 —— 筛选细分访客，页面优化有的放矢

企业官网是企业潜在客户的指路牌。某 2B 企业官网运营人员，根据用户的

官网访问时长、用户行为路径、活跃度、注册与否等因素，将用户细分为单纯浏览者、信息收集者和购买需求强烈者三类。

运营人员事先按照自定义规则，将三类访客进行用户分群。接下来，在“点击分析”功能模块中，分别筛选出三类人群，并查看其页面点击情况。

1. 用户群体之“单纯浏览者”的点击分析与优化方法。

如图3-27所示，该类用户群体对“产品介绍”、“视频”点击率较高，这说明“产品介绍”、“视频”是初来乍到的访问者了解企业的“窗口”，而元素内容缺少“亮点”，导致访问者不愿意花时间停留。

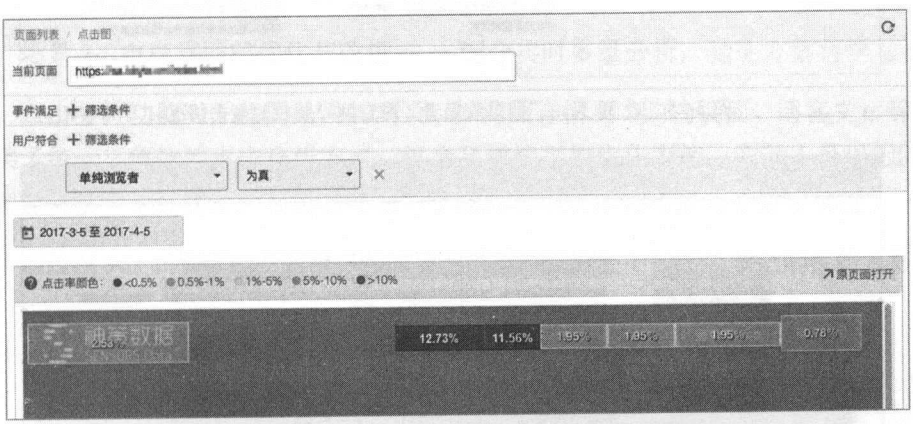


图3-27　近30天，“单纯浏览者”对官网导航栏的点击情况

因此我们可尝试以下两方面的优化。一方面，优化内容。让产品价值、优势、案例等内容尽可能醒目，以快速吸引浏览者注意；另一方面，在导航栏中尝试增加社交因素。如建立论坛、设立产品博客，以增强访问者对官网的黏性，提高网站的活跃用户数量。

2. 用户群体之“信息收集者”的点击分析与优化方向。

官网运营人员应该帮助该用户群体确定购买意向。图3-28显示，“信息收集者”群体对官网导航条中“文档”、“博客”兴趣很高，而行业解决方案的点击较少。事实上，行业解决方案是该类群体值得关注的价值点，由于点击较低，可以尝试将其调整至醒目位置，进行效果对比。

在点击率颇高的“文档”栏目中，哪种类型文章最受钟爱？运营人员可以直接点击“文档”自动进入“文档”的点击分析页面，如图3-29所示，点击数据表

明，技术剖析、案例精选、业务场景分析的点击率最高，运营人员可持续优化群体关心的“产品”、“博客”等展现形式和内容。

图 3-28　近 30 天，“信息收集者”对官网导航栏的点击情况

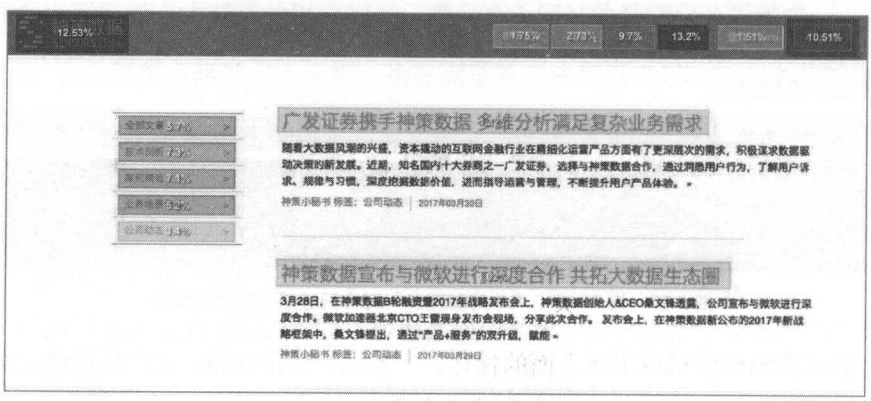


图 3-29　点击“文档”，直接跳转文档页面并展示点击情况

3．用户群体之“需求强烈者”的点击分析与优化方向。

值得一提的是，神策分析的“点击分析”功能支持弹框的点击统计。按照最初用户分群的定义，此群体都点击了“申请试用”按钮，而图 3-30 显示只有 12.11% 的用户最终提交。因此设计人员该反思，是否要填的注册信息过多导致用户点击注册而放弃提交。为了提升用户的试用体验，还可尝试将帮助中心、在线客服等帮助链接附在显著位置，使用户在整个试用过程中能随时得到帮助。

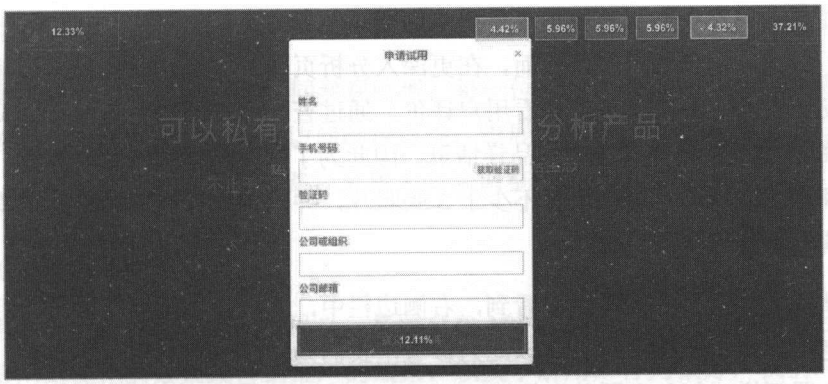


图 3-30　点击“申请试用”按钮群体的点击分析

场景 2：电商界面的优化与改进——配合实时多维分析，验证方案科学与否

点击分析功能对于相同结构的网页，如商品详情页、购物页面、博客文章等，提供了统一、便捷的点击分析方式。以商品详情页的优化为例，产品人员以 URL 规则建立了一个页面组，并选择任意一个商品详情页作为背景展示点击情况。

通过图 3-31 我们可以看到用户在该页面频繁地点击商品的图片和已购买的人数。

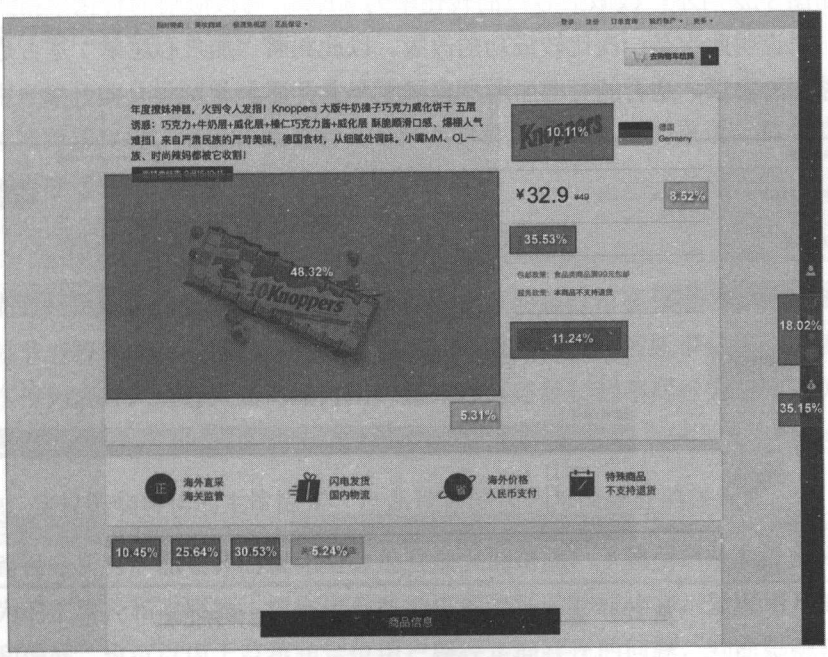


图 3-31　该网站中商品详情页的点击图情况

显然，用户在购买前希望了解更多的商品信息，尤其是图片、已购买用户的评价，进而决定是否下单。然而，在更深入分析页面时发现，商品图片只有 1 张且不支持查看大图，又无法查看用户评价。通过查看网站的历史数据，每天大约有 50% 的用户来浏览这样的商品详情页。因此为了优化目标页的用户体验，可以要求商家发布商品时必须上传不少于 3 张照片，并支持所有类型的商品详情页都有已购买者的评价。

从商品详情页的点击图中看到，右侧边栏中，“我的心愿单”按钮被用户，尤其老用户点击的频率很高。以此为参考，我们为页面改版找到一些方向：在合适的位置新增“加入心愿单”按钮。

改版后，产品人员再次通过点击分析工具评估效果时发现，“加入心愿单”按钮的点击率达到 30%，而“立即购买”按钮的点击率只下降了 1%。说明这次改版对“立即购买”按钮的点击率的冲击程度不大，并不会影响页面的最终转化。

“加入心愿单”是否对用户转化造成影响？产品人员可通过分布分析“加入心愿单”操作的频率和人数，或者通过留存率判断用户黏性的强弱变化。

如图 3-32 所示，改版后客户的转化率为 3.17%，与改版前的转化率相比，若变高，则说明此次是一次比较成功的改版。以此判断“加入心愿单”是否是用户真实存在的需求，是否能对增加用户忠诚度产生贡献。

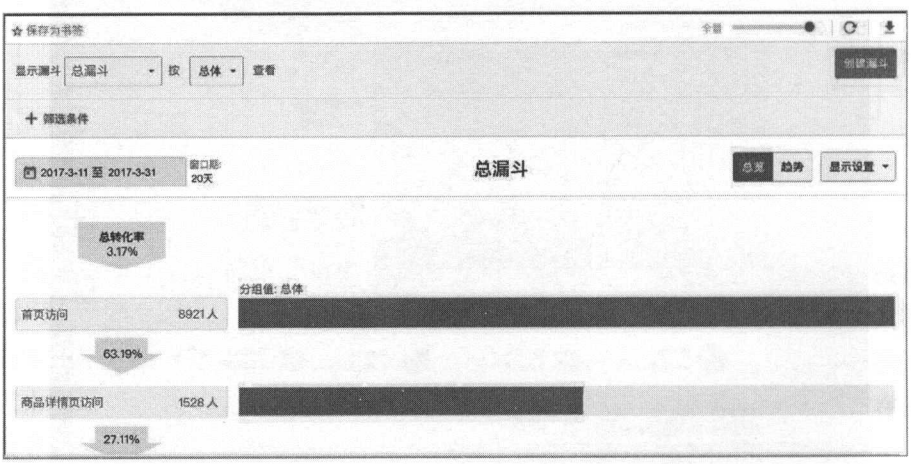


图 3-32　通过漏斗分析功能，查看改版后的总体转化率

用户路径

用户访问 APP 或网站，如同参观画展，每一位观众根据自身喜好形成特有的参观顺序。为让观众沿着最优访问路径前进，策展者需要结合观众需求进行布局调整。从一定程度上讲，用户路径分析为布局调整提供科学指导。

用户行为路径的定义

用户行为路径分析，顾名思义，是用户在 APP 或网站中的访问行为路径。为了衡量网站优化的效果或营销推广的效果，以及了解用户行为偏好，我们时常要对访问路径的转换数据进行分析。

以电商为例，买家从登录到支付成功要经过首页浏览、搜索商品、加入购物车、提交订单、支付订单等过程。而用户真实的选购过程是一个交缠反复的过程，例如提交订单后，用户可能会返回首页继续搜索商品，也可能去取消订单，每一个路径背后都有不同的动机。与其他分析模型配合进行深入分析后，找到快速用户动机，从而引领用户走向最优路径或者期望中的路径。

用户路径分析模型的价值

用户路径的分析结果通常以桑基图形式展现，以目标事件为起点或终点，查看后续或前置路径，可以详细查看某个节点事件的流向，总的来说，科学的用户路径分析能够带来以下价值。

1. 可视化用户流，全面了解用户整体行为路径。

通过用户路径分析，可以将一个事件的上下游进行可视化展示。用户即可查看当前节点事件的相关信息，包括事件名、分组属性值、后续事件统计、流失、后续事件列表等。运营人员可通过用户整体行为路径找到不同行为之间的关系，挖掘规律并找到瓶颈。

2. 定位影响转化的主次因素，产品设计的优化与改进有的放矢。

路径分析对产品设计的优化与改进有很大的帮助，了解用户从登录到购买整体行为的主路径和次路径，根据用户路径中各个环节的转化率，发现用户的行为规律和偏好，也可以用于监测和定位用户路径走向存在的问题，判断影响转化的主要因素和次要因素，并发现某些冷僻的功能点。

用户路径的应用场景

场景：启动 APP 后，为何只有 30% 商超客户交易成功

这是全球领先的社区 O2O 服务平台中商惠民[1]的数据分析场景。

在一次评估客户总体转化率过程中，运营人员通过漏斗分析发现，从登录惠配通 APP 后，提交订单的商超客户仅有 30%，接下来运营人员通过用户路径分析客户流失的原因所在，用户路径分析模型清晰展示了商超客户的动作走向。

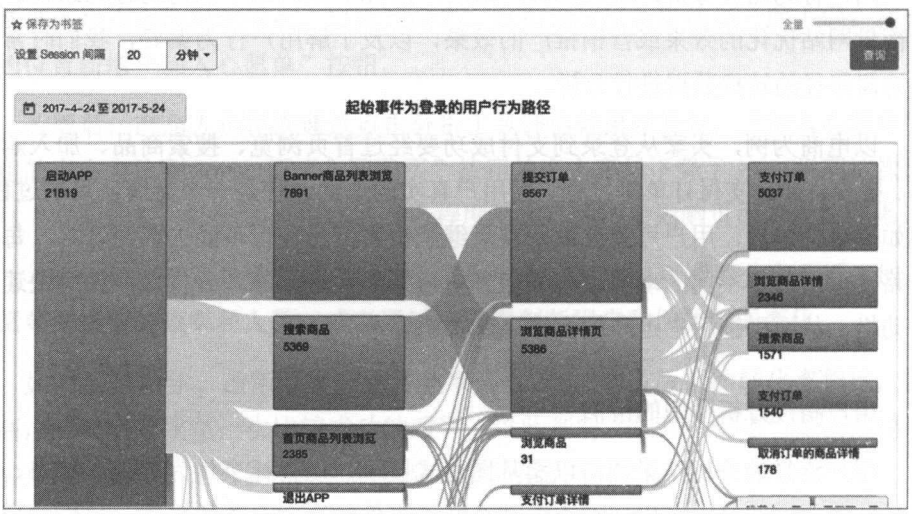


图 3-33　用户行为路径分析

中商惠民运营人员选取若干事件对客户购买路径进行深度分析。图 3-33 显示，用户登录 APP 后，约有 40% 的客户会点击 Banner，约 30% 的客户会直接进行商品搜索，约 10% 的用户会浏览商品列表，约 5% 的客户直接退出 APP。

运营人员进一步查看这 4 类用户的提交订单的情况，直接进行“搜索商品”的用户进行提交订单比例最高，超过 90%；与其形成鲜明对比的是，尽管“点击 Banner”是更多客户登录 APP 后的首选动作（约占总客户的 40%），但这部分用户群体在浏览商品列表后，仅 30% 的用户提交订单，说明 Banner 内容布局有比较糟糕的用户体验，因此这就成为企业首选优化与改进的方向。

[1] 因涉嫌商业机密，所涉数据均为虚拟。

用户分群

通过漏斗分析可以看到，用户在不同阶段所表现出的行为是不同的，譬如新用户的关注点在哪里？已购用户什么情况下会再次付费？因为群体特征不同，行为会有很大差别，因此可以根据历史数据将用户进行划分，进而再次观察该群体的具体行为。这就是用户分群的原理。

用户分群的定义

用户分群即用户信息标签化，通过用户的历史行为路径、行为特征、偏好等属性，我们将具有相同属性的用户划分为一个群体，并进行后续分析。

用户分群分析的分类与价值

用户分群通常被分为普通分群和预测分群。普通分群根据用户的属性特征和行为特征将用户群体进行分类，预测分群根据用户以往的行为属性特征，运用机器学习算法来预测他们将来会发生某些事件的概率。

举例来说，互联网金融产品的用户按照风险投资偏好这一属性分为保守型、稳健型和激进型，按照投资行为分为已投资和未投资。运营人员可以根据这一属性和行为将满足某种条件的用户群体提取出来，譬如激进型但未投资的这群用户，分析这一群体的行为特征从而优化产品促进用户投资，或者根据其浏览的项目页面，推荐用户可能会感兴趣的项目。

总的来说，用户分群具有以下价值。

1. 帮助企业打破数据孤岛并真实了解用户。

用户画像是用户分群的前提，对特定属性的用户群体进行持续深入的用户行为的洞察后，该用户群体的画像逐渐清晰。这些都有助于企业了解某个指标数字背后的用户群体具备哪些特征 —— 他们是谁？行为特点有哪些？偏好是什么？潜在需求和行为喜好是什么？这是后续用户群体针对性分析的前提。

2. 定位营销目标群体，帮助企业实现精准、高效营销。

清晰勾勒某群体在特定研究范围内的行为全貌，并定义目标人群，是运营人员信息推送的前提。运营人员根据需求对特定目标人群完成精准信息推送工作，如召回流失用户、刺激用户复购等。当完成特定人群的精准信息推送工作之后，

进一步分析以实时全方位查看营销效果，帮助企业与用户实现精准高效的信息互通。

图 3-34 显示了高黏性与高频消费用户的筛选。

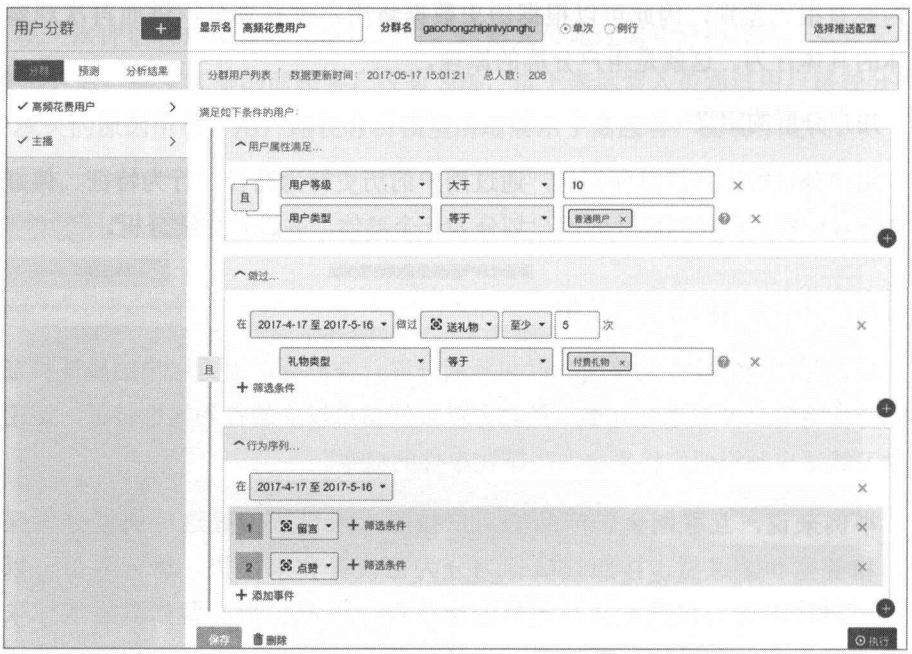


图 3-34　高黏性与高频消费用户的筛选

用户分群应用场景

场景 1：直播行业 —— 高黏性与高频消费用户的行为观察

以直播产品为例，运营人员可以筛选出过去 30 天内、等级 10 级以上、有“留言”和“点赞”行为，并且付费礼物送出次数超过 10 次的用户，我们视其为高黏性且高频消费用户，并将其进行分群定义。

图 3-35 是高黏性与高频消费用户列表明细。

运营人员可通过事件分析来观察这部分用户群体近期的行为表现，图 3-36 显示该用户群体的人均观看时长与其他用户存在明显差别。

用户分群 / 用户列表　　列目显示

distinct_id	广告系列媒介	广告系列来源	最近充值时间	最近花费时间	最近访问时间
003ddea8ba3fdae0	头条_iOS	今日头条	2017-01-15 13:03:49.090	2017-01-15 13:03:49.090	2017-01-15 13:03:49.090
00f437957f236b8a	微信_iOS	今日头条	2017-01-02 19:08:34.285	2017-01-02 19:08:34.285	2017-01-02 19:08:34.285
0128df11f05d40f2	微信_iOS	今日头条	2017-02-14 22:26:22.165	2017-02-14 22:26:22.165	2017-02-14 22:26:22.165
015509dfab2a19b2	微信_iOS	36kr	2017-01-04 18:50:20.688	2017-01-04 18:50:20.688	2017-01-04 18:50:20.688
01b85bade4a26aa2	微信_iOS	百度	2017-01-29 17:02:54.193	2017-01-29 17:02:54.193	2017-01-29 17:02:54.193
02a09c0a0af2a93e	微信_iOS	百度	2017-01-03 10:12:43.636	2017-01-03 10:12:43.636	2017-01-03 10:12:43.636
058b5b1881b47fd1	头条_Android	百度	2017-01-08 09:17:19.436	2017-01-08 09:17:19.436	2017-01-08 09:17:19.436
064bbbf768d01fba	微信_iOS	36kr	2017-04-07 14:54:55.773	2017-04-07 14:54:55.773	2017-04-07 14:54:55.773
07dbf7f72a697716	微信_iOS	百度	2017-01-05 07:01:54.285	2017-01-05 07:01:54.285	2017-01-05 07:01:54.285
07f358816c9bf5d8	头条_iOS	36kr	2017-05-09 13:22:54.545	2017-05-09 13:22:54.545	2017-05-09 13:22:54.545
08da366ff528882f	微信_Android	地推	2017-01-07 13:20:39.999	2017-01-07 13:20:39.999	2017-01-07 13:20:39.999
0901c337c6f8027c	头条_iOS	今日头条	2017-01-10 10:18:09.075	2017-01-10 10:18:09.075	2017-01-10 10:18:09.075
0ae134ea753ea108	头条_iOS	百度	2017-04-14 10:27:08.958	2017-04-14 10:27:08.958	2017-04-14 10:27:08.958
0b3fbcffe84ebbfb	微信_Android	36kr	2017-01-18 18:40:14.116	2017-01-18 18:40:14.116	2017-01-18 18:40:14.116
0c9960695b4e711f	微信_Android	地推	2017-01-02 20:33:16.363	2017-01-02 20:33:16.363	2017-01-02 20:33:16.363
0d9bc96956359062	微信_iOS	今日头条	2017-01-19 02:27:35.999	2017-01-19 02:27:35.999	2017-01-19 02:27:35.999
0dad1f1b91c39e7a	微信_iOS	36kr	2017-01-14 16:47:19.023	2017-01-14 16:47:19.023	2017-01-14 16:47:19.023
0deafe17ff8b7670	头条_iOS	百度	2017-01-20 06:06:49.090	2017-01-20 06:06:49.090	2017-01-20 06:06:49.090
0fae7ee9506c4aa2	微信_iOS	百度	2017-01-21 14:46:12.413	2017-01-21 14:46:12.413	2017-01-21 14:46:12.413
0fc0fbba73a5a8d1	头条_Android	百度	2017-02-23 13:17:32.307	2017-02-23 13:17:32.307	2017-02-23 13:17:32.307

图 3-35　高黏性与高频消费用户列表明细查询

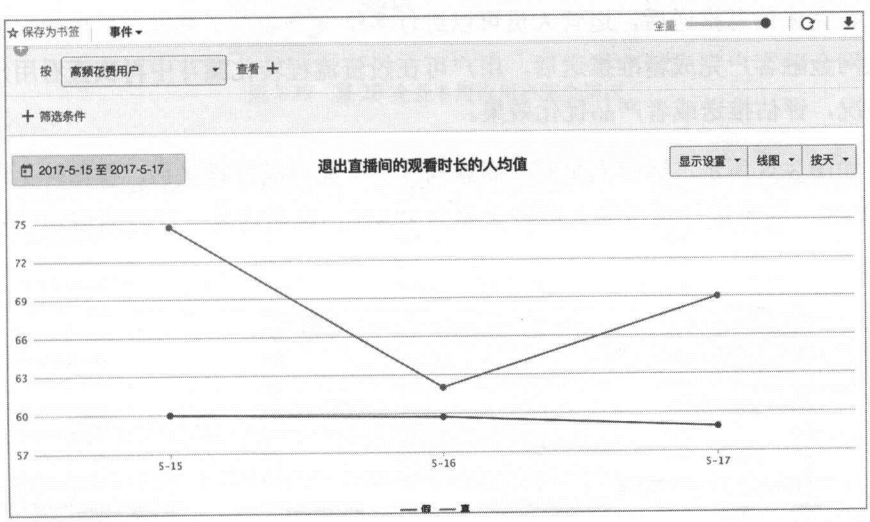


图 3-36　高频花费用户与非高频花费用户观看时长人均值对比

场景 2：精准营销及效果评估案例分享

1．互联网金融行业——唤醒“沉睡”用户的精准推送与效果评估。

例如，某互联网金融客户为“唤醒”2017 年 1 月注册且浏览过征信页面（通

过分析发现，用户浏览征信页面后，后期的留存率较高），但未进行投资的用户，并向该群体推送“将于 1 月 20 日起发行贺岁版理财，预期年化收益率高达 9.50%”的信息。为锁定目标人群，运营人员在用户分析模块的“用户分群”功能页面做如图 3-37 所示的操作。

图 3-37　在“用户分群”功能页面，筛选营销目标群体

在完成信息推送后，运营人员可以进行多维度分析，了解推送后效果。如该互联网金融客户完成精准推送后，用户可在投资流程转化漏斗中再次查看用户转化情况，评估推送或者产品优化效果。

如图 3-38 所示，运营人员对“高意向客户”完成精准推送后，整体转化率高达 78.26%，而未进行推送的人群转化率为 77.83%，说明这是一次较为成功的精准营销。

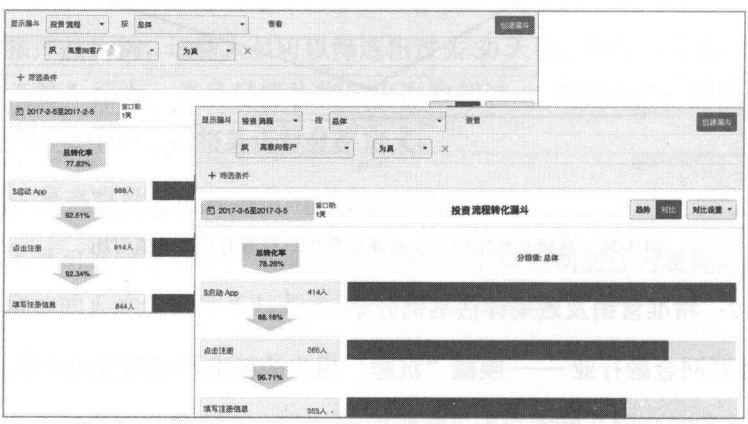


图 3-38　被推送人群与未被推送人群的总体转化率情况对比

2. 企业级服务（2B）——“召回”流失客户的精准推送和效果评估。

某 2B 企业以投资到期之后再次投资作为留存的标准，近 8 周用户流失情况如图 3-39 所示。在完成筛选工作后，运营人员可在用户明细页面上，直接将该用户群体进行定义，在此基础上完成精细化推送工作。

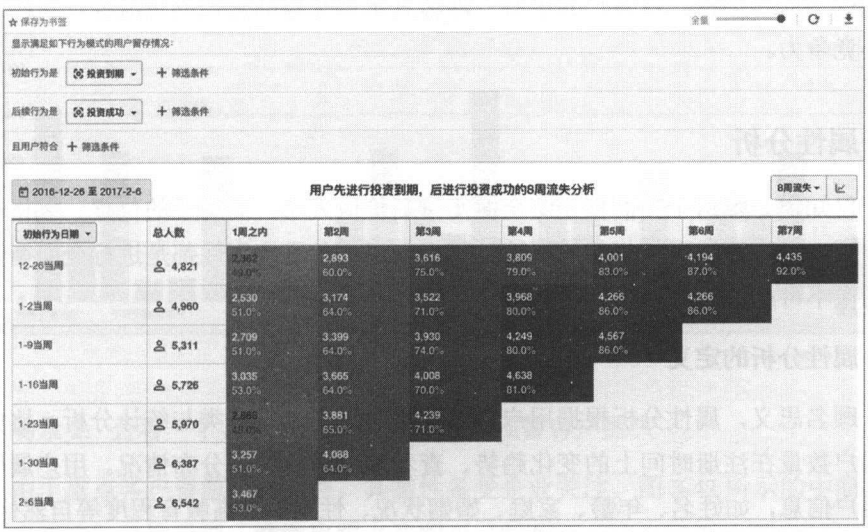


图 3-39　某 2B 企业 8 周内用户流失情况

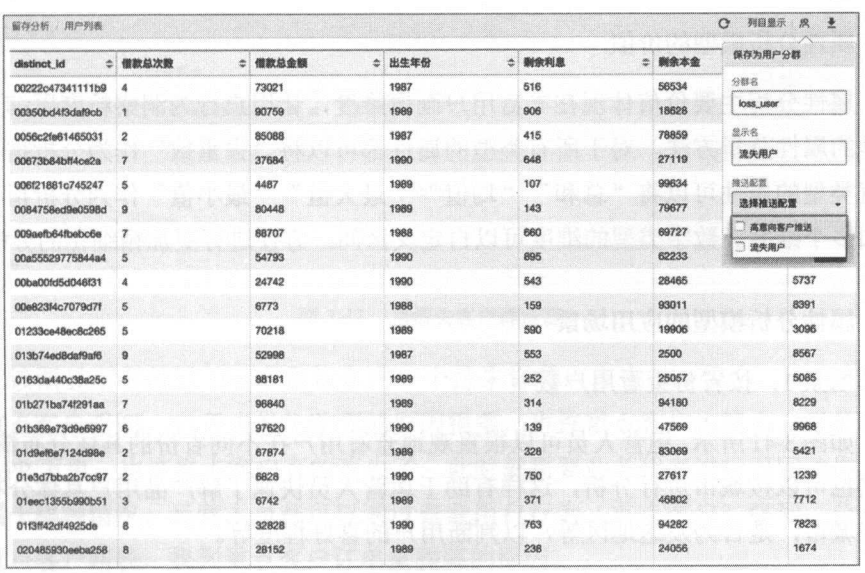


图 3-40　查看特定时间内的流失人群明细，并保存为用户分群，命显示名为“流失用户”

在该页面上，运营人员点击留存数值，即查看流失人群的详细信息，并直接创建用户分群并推送信息，以刺激其申请产品使用，如图 3-40 所示。

大数据时代，为适应不断变化的外部市场环境，提升客户黏性，企业不断加速数字化营销转型。其中，提升营销效率、提高营销精准度是企业的首要战略目标。以上场景都将“以客户为中心”理念真正贯穿精准营销的全流程，重构企业核心竞争力。

属性分析

仅知道一幢房子的面积无法全面衡量其价值大小，而房子的位置、风格、是否学区、交通环境更是相关的属性。同样，用户各维度属性都是进行全面衡量用户画像不可或缺的内容。

属性分析的定义

顾名思义，属性分析根据用户自身属性对用户进行分类与统计分析，比如查看用户数量在注册时间上的变化趋势、查看用户按省份的分布情况。用户属性涉及用户信息，如姓名、年龄、家庭、婚姻状况、性别、最高教育程度等自然信息，也有产品相关属性，如用户常驻省市、用户等级、用户首次访问渠道来源等。

属性分析模型的价值

属性分析主要价值体现在丰富用户画像维度，让用户行为洞察粒度更细致。科学的属性分析方法，对于所有类型的属性都可以将“去重数”作为分析指标，数值类型的属性可以将“总和”“均值”“最大值”“最小值”作为分析指标，添加多个维度。数字类型的维度可以自定义区间，方便进行更加精细化的分析。

属性分析模型的应用场景

场景 1：按省份查看用户数

如图3-41所示，运营人员可以很直观地查看用户在不同省份的具体分布情况。当然也可以按城市进行分析，这样有助于运营人员快速了解产品用户具体分布在哪些城市，是否为发达地区等，以判断用户的喜好程度等。

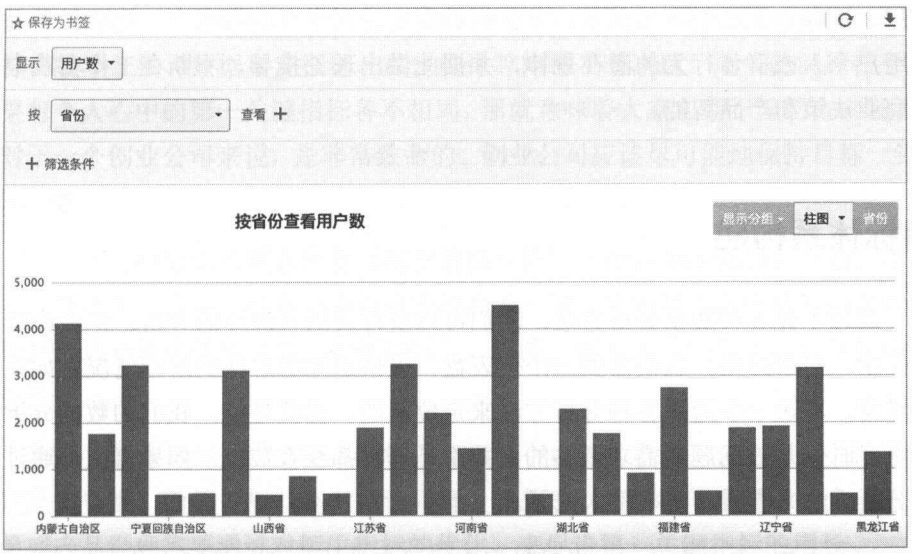


图 3-41　按省份查看用户数

场景 2：查看一个月未发生购买的客户，预警客户流失

由于重点客户资源的稀缺性，其黏性备受企业关注。图 3-42 所示的中商惠民“用户属性”分析模型，筛选出距上次购买已经超过一个月的重点客户。

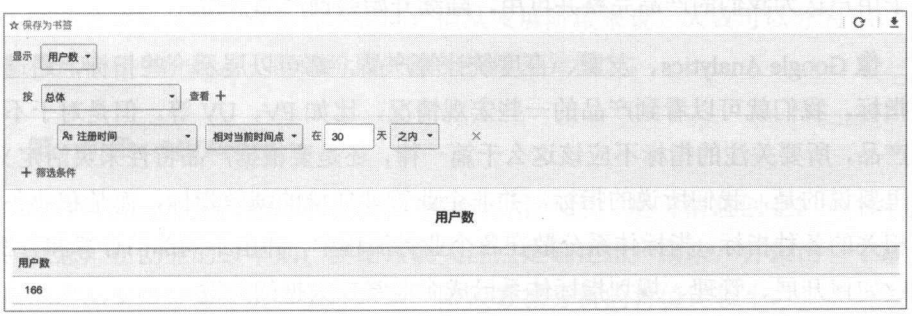


图 3-42　准流失客户群预警

该图显示有超过 166 个商超客户一个月未订货。点击 166 数字，即呈现 166 家重点客户明细。30 天未发生购买的原因很多，也许重点客户未流失，只是不再用 APP 下单，也许重点客户真的流失了。此时就需要业务代表进行召回动作，无论属于哪种情况，运营人员都可以通过查看用户行为（重点客户）序列，分别了解重点客户路径，找到重点客户订单量骤降的原因。

我们逐一介绍了各大数据分析模型，实际上各分析模型应该是一个综合体，

它们之间存在无法割裂的关系。各分析模型实现科学互动和配合，能够科学揭示出用户个人或群体行为的潜在规律，并据此做出理论推导，不断在工作实践中优化商业决策和产品智能。

指标体系构建

通过上述 8 种数据分析方法，我们可以进行灵活且深度的分析。但在实际的工作中，更多时候，我们需要一个仪表盘，以快速知道总体的运营情况。如果发现异常，再进一步通过各种分析方法来定位问题。也就是说，我们的数据分析方法更多时候是被问题或需求驱动的，并不是每次都要看数据。如果每次都通过这些分析方法来查看，效率就太低了。

创业之初，我们花了一个半月时间完成了产品的 0.1 版，该版本共有三个功能，分别是事件分析、漏斗分析和留存分析。我们将 Demo 产品拿给种子用户体验，种子用户体验后觉得很不错，认为这种灵活的多维分析能力很强大，但是缺少一个功能——数据概览，即将常用的指标直接配置好，方便查看。于是我们又花了一个月时间，专门做了“数据概览”功能，并且支持在手机浏览器上展示。这时种子用户认为我们的产品完整并可用，陆续开始咨询产品价格。

像 Google Analytics、友盟、百度统计等产品，都可以展示一些指标，通过这些指标，我们就可以看到产品的一些宏观情况，比如 PV、UV 等。但是对于不同的产品，所要关注的指标不应该这么千篇一律，还是要根据产品特性来灵活定义。这里要说的是，我们所说的指标，并非企业管理环境的绩效指标，而是和业务运营相关的各种指标。指标体系分散在各个业务流程中，并由不同部门计算和分析。企业如何开展、管理、规划指标体系已成企业掌控数据的关键。

但许多互联网初创公司，以及一些互联网 + 类的公司，并不清楚该如何结合自己的产品特征，定义合理的指标体系。接下来，我给大家介绍两套构建指标体系的方法，分别是第一关键指标法和海盗指标法。

第一关键指标法

对你所负责的产品来说，最关键的指标是什么？

经过调研，我们发现每家企业的不同员工对第一关键指标认知不一样，甚至

有些员工对第一关键指标没有概念。同时，伴随企业各部门应用的指标越来越多，指标口径不一致、数据重复报送、重复统计等问题不断涌现。更大的问题在于，如果每个人心中的第一关键指标各不相同，那就意味着大家在朝不同的方向发力，这对于一个创业公司来说，是非常致命的。创业公司应该尽可能地保持目标一致、行动一致。

第一关键指标法的概念出自《精益数据分析》（*Lean Analytics*）一书，即在企业发展的每个阶段，都有一个当前阶段高于一切、需要集中全部精力注意的一个数据，这个数据就是“第一关键指标”。当然，随着业务的发展，这个指标会发生变化。

第一关键指标可外延出更多指标，比如一个成熟的电商平台，第一关键指标一定是销售额，而销售额能够衍生访问量、转化率、客单价等多指标。企业运营人员或产品经理需要通过对衍生指标的优化，来促进第一关键指标的增长。

因此，企业应该基于第一关键指标及衍生指标来衡量发展情况。让全企业员工明确当前阶段的核心目标，以此来制定与规划清晰任务。第一关键指标法和绩效管理中的 KPI 的理念比较接近，就是要寻找当前阶段整个公司最需要关注的指标，以此来集中火力向目标前进。

创业公司所从事的业务各不相同，但从发展阶段来说，大致可以分为 MVP、增长和营收三个阶段，不同的阶段关注的指标差异很大。

第一阶段：MVP 阶段

MVP（Minimium Viable Product，最小可用产品）是《精益创业》一书中提出的理念，在创业的最早期，创始人的关注点是确定用户需求，并做出一个最小可用的产品来验证需求的真实性。这一阶段数据分析的价值比较小，企业需要定性分析，如通过大量的用户访谈来确定产品的满足情况，此阶段并不需要在数据分析方面投入大量工作。

第二阶段：增长阶段

此阶段的企业已经有了成型的产品以及固定的用户群，有丰富的数据可以进行数据分析。我们将数据分析细分为关注留存指标和关注引荐指标两个阶段。

在产品推广之前，留存分析直接反映了用户的活跃度，帮助企业证明产品是

否给客户带来价值，同时，产品经过不断优化和迭代，企业应该关注引荐指标。我的知乎系列文章《PayPal[1] 早期是如何通过 Growth Hacking 成长为独角兽的》中介绍到，PayPal 用户向朋友转钱，会促使该朋友也注册 PayPal。企业也可以人为促进这种传播速度，比如 PayPal 每引荐一位朋友注册，就可得到 10 美元的奖励。在这个阶段中，病毒系数和病毒周期值得关注，假如在一年的时间内，1 个用户推荐 2 个用户成功注册，那么病毒系数就是 2，病毒周期就是 0.5 年。

当然并非每个产品都能找到病毒传播的途径，若不能，也可尝试口碑的力量。神策数据就是用 NPS（Net Promoter Score，净推荐值）来评估客户满意度的：企业级服务的本质是为客户创造价值，客户发展应像“滚雪球”一样越滚越大，在精心培养老客户忠诚度的基础上，再开拓新客户。而低劣的产品与服务只能造成“猴子掰玉米”般的客户流失。NPS 在实际经营中有可能粒度太粗，实际的经营活动作用到 NPS 的周期可能太长。为此，我们又创建了衍生指标，即每个企业平均每周每账户的有效查询次数。通过这一指标，我们可以监控客户的使用情况，如果太低的话，就要和客户沟通到底是在数据采集还是在指标配置上出了问题。只要提升了有效查询频次，就可以提升 NPS。

第三阶段：营收阶段

此阶段产品形态已相对成熟，企业的关注点聚焦在如何规模化，并实现快速盈利，关键指标主要是 LTV（Life Time Value，生命周期总价值）、CAC（Customer Acquisition Cost，用户获取成本）、渠道分成比例、渠道用户盈利周期等。此时企业需要寻找新的发展方向，为下一步增长做准备，而新方向则可以重复这三个阶段。

以上是不同阶段的典型关键指标，总之，第一关键指标应该具备能够正确反映业务和阶段、简单易懂、具有指导性的特征。对于项目本身还应该具体问题具体分析，比如一直处于增长阶段的百度知道，产品已非常程度成熟，属于百度搜索的子产品，它无需过于关注盈利的问题。

案例：为什么将“提升回答量”定为第一关键指标

在本书的第 1 章中，我提到在百度知道提升回答量的实践。在着手这项工作

[1] PayPal，全球众多用户使用的国际贸易支付工具。1998 年 12 月由 Peter Thiel 及 Max Levchin 建立，总部位于美国加利福尼亚州圣荷西市。自 2002 年出售给 eBay 之后，PayPal 的大部分重要员工都已经离职，但他们仍然保持着密切的联系。PayPal 堪称是创业者的摇篮，他们被誉为“PayPal 黑帮”。

之前，我们经历较长时间的摸索才定下“提升回答量”为第一关键指标。

我加入百度后，才知道要做后端研发，我每天会收到包含各类统计数据的报表邮件，这些数据包括百度知道的访问量、检索量、独立 IP 数、Session 数、提问量、回答量、设置最佳答案的数量等一系列指标。当时我并无“第一关键指标”的认知，觉得什么指标都很重要，后来我才明白产品优化不能胡子眉毛一把抓，一定要有的放矢。

最初我排除了访问量、检索量这些指标，因为这些指标渠道影响很大，即百度大搜索。相比之下，提问量和回答量更重要，而双指标发力让我有心无力。

在一次产品沟通会上，我向当时的产品经理孙云丰请教了这个问题。他告诉我，提问量并非关键问题，因为用户在百度搜索过程中，若发现没有很好的答案时候，只要加个引导链接，就能导流过来大量提问。我顿时醍醐灌顶：第一关键指标是回答量，之后我所做的事情则围绕提升回答量展开，特别是做了问题推荐项目，以提升回答量。

不只是企业，即使对于国家而言，第一关键指标也可以带来巨大的价值。我们国家从 1978 年改革开放以来，确定以经济建设为中心。为此，政府采用 GDP（国内生产总值）作为第一关键指标。几十年来，经济取得了突飞猛进的成绩，如图 3-43 所示。

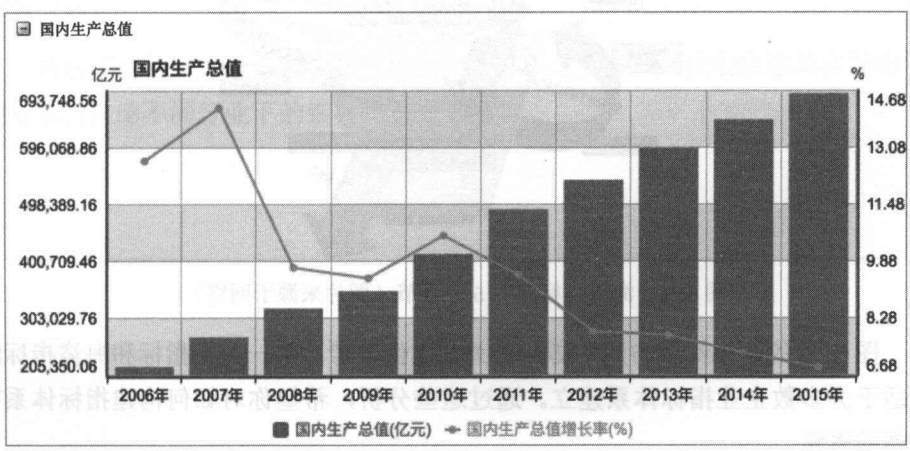


图 3-43　1978—2015 年，我国 GDP 的发展趋势

总之，若企业同时对多指标进行跟进与优化，就会导致团队专注度下降，让整个公司拧成一股绳朝着相同的指标前进才是最佳方式。

海盗指标法

第一关键指标法简单有效，但对于实际的产品运营来说，我们需要全方位地做出监测，这时海盗指标法就派上用场了。

2007 年，500 Startups 创业孵化器的创始合伙人戴夫 • 麦克卢尔（Dave McClure）针对创业公司应该关注的指标，提出了一套模型——Pirate Metrics，即海盗指标法。他将创业公司需要关注的指标归结为 5 个方面，分别是 Acquisition（获取）、Activation（激活）、Retention（留存）、Revenue（营收）和 Referral（引荐），简称 AARRR。

图 3-44 介绍了海盗指标法对应用户生命周期中的 5 个重要环节。每一环节都有需要衡量的指标。相比第一关键指标法，海盗指标法的阶段划分更加清晰，但因戴夫 • 麦克卢尔是营销出身，因此侧重点在营销方面。由于 2007 年他在介绍此模型时，还没有出现移动互联网市场，所以图片上并没有显示 APP 类的一些业务场景。第 4 章还会进一步讲解这 5 个环节。

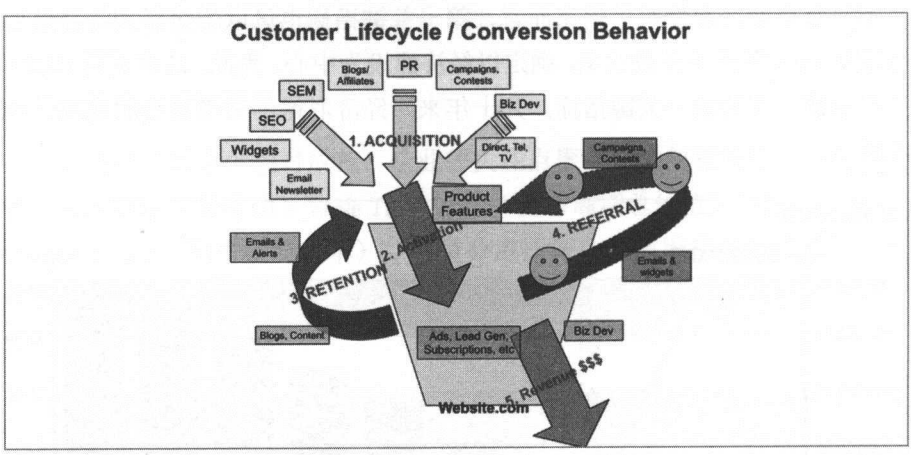


图 3-44　海盗指标法的 5 个环节（图片来源于网络）

图 3-45 的各指标都与网站访问相关，仅供参考。第一关键指标和海盗指标法普适于大多数企业指标体系建立。通过这些分析，希望你对如何构建指标体系有清晰的思路。

Category	User Status	Conv %	Est. Value
Acquisition	**Visit Site** (or landing page, or external widget)	100%	$.01
Acquisition	**Doesn't Abandon** (views 2+ pages, stays 10+ sec, 2+ clicks)	70%	$.05
Activation	**Happy 1st Visit** (views X pages, stays Y sec, Z clicks)	30%	$.25
Activation	**Email/Blog/RSS/Widget Signup** (anything that could lead to repeat visit)	5%	$1
Activation	**Acct Signup** (includes profile data)	2%	$3
Retention	**Email Open / RSS view -> Clickthru**	3%	$2
Retention	**Repeat Visitor** (3+ visits in first 30 days)	2%	$5
Referral	**Refer 1+ users who visit site**	2%	$3
Referral	**Refer 1+ users who activate**	1%	$10
Revenue	**User generates minimum revenue**	2%	$5
Revenue	**User generates break-even revenue**	1%	$25

图 3-45　海盗指标法指标设计（图片来源于网络）

第一关键指标法定位了企业当前发展阶段的最重要问题，它关注全企业层面的运转健康，有利于让全公司形成合力聚焦同一目标。

海盗指标法为企业提供了数据分析基础和罗盘，以及指导创业和企业发展的探索方向。

然而不同行业、不同的商业模式的指标体系差异较大，我们应该从实际出发。第 6 章将介绍不同行业下的指标体系差异性。

第 4 章

数据驱动产品和运营决策

在第2章中，我们介绍了数据驱动“决策”和数据驱动“产品智能”两个方面。第4章与第5章将围绕这两个方面进一步介绍。

所谓数据驱动决策，就是通过数据来指导人做决定。在互联网产品中，决策包括运营监控、产品改进和商业决策三个方面。

数据驱动运营监控

数据驱动运营监控，即企业管理者和运营人员能够密切关注企业市场运营、用户运营、内容运营、社区运营以及商务运营等环节的运行实况，通过日常对一些数据的观察，企业可以客观评价、优化、改进运行流程与问题。

在前面提到由戴夫·麦克卢尔针对创业公司提出的海盗模型中，他将创业公司需要关注的指标归结为5个方面：Acquisition（触达）、Activation（激活）、Retention（留存）、Referral（引荐）、Revenue（营收），简称AARRR。这5个方面都会涉及大量的运营工作，下面我对此部分进行深入讲解。

用户获取（Acquisition）

“我知道我一半的广告预算都浪费了，只是不知道是哪一半。”美国百货之父约翰·华纳梅克（John Wanamaker）曾这样说。企业市场部通过不同渠道触达用户，好的营销渠道带来的用户与设定的目标人群有很大吻合度，有针对性地圈定了目标人群，坏的营销渠道则相反。作为“花钱大户”的市场部门来说，衡量各渠道的ROI是重中之重。尤其对于初创企业来说，盲目试错造成的损失巨大。

只有甄选出最优渠道，才能实现营销资源和营销渠道的把控。图 4-1 显示了不同渠道来源的触发用户数。

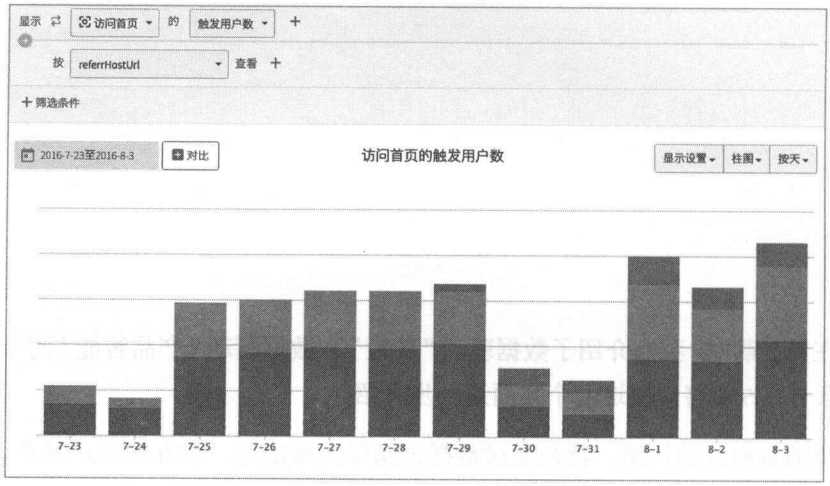


图 4-1　不同渠道来源的触发用户数

相较于线下传统营销渠道，数字营销渠道如官网、微信公众号、EDM 营销、线下活动等都可以被跟踪和监测。通过用户行为数据分析，运营人员可以科学评估数字营销各渠道的投入产出比，好让市场管理者能够进行渠道追踪，实时把握渠道动态，对市场活动有效性加以计划、执行、监控和分析。

乐纯案例：全渠道市场营销管理，优化营销投入产出比

乐纯酸奶是一家创新型的食品公司，营销分成线上和线下。线下业务人员主要负责地推工作，相当于一支直销团队，由业务人员直接接触终端消费者后进行活动与试吃。传统渠道推广的数据非常零碎且割裂，然而乐纯通过数据分析系统打通了线上线下的数据接入。

乐纯为每位地推人员分配了不同的二维码，这样就可以跟踪每位地推人员的实际推广效果。消费者扫描二维码进入产品，系统会自动记录二维码来源，映射到对应的地推人员，客户在产品页面进行的操作则全部记录。同时，如微信“阅读原文”、不同广告渠道来源数据也都被记录下来，以清晰分析哪些渠道更加有效。

激活（Activation）

什么是激活？我们不妨从“激活”常见的认知误区开始介绍。

激活的认知两大误区

误区 1：激活意味着用户触达

用户触达是用户激活的前提。企业采用 PR、广告投放等手段，目的是让用户先知道产品，然后再去尝试。

然而，单纯的用户触达没有直接对后续转化带来价值，不能保证留存率。既然企业为每条线索都付出了高额成本，用户从不同的渠道“认识”产品时，若只是启动 APP 或者打开网页，但未注册、购买就离开，就会造成前期工作功亏一篑。

因此一些企业将用户下载并启动一次 APP 定义为一个有效的激活用户，显然是一个误区。再如，对于 SEM 投放来说，搜索引擎公司帮助企业解决了用户触达的问题，并按照链接被点击的次数来收取费用——当用户被引导到企业网站，企业就要付费。

归根到底，我们在衡量市场活动效果时，需要以用户激活量来衡量，而非用户触达量。除非你只是为了进行品牌宣传。对于一家初创公司来说，效果类活动应该优先于品牌宣传类活动，毕竟弹药有限。

误区 2：激活味意着注册成功

新增用户数是产品所关注的用户指标之一。注册一个新的用户 ID 即新增一个用户数。那么，成功注册能代表激活用户吗？不能。

我们所说的用户激活，是指新用户真正体验了产品的核心功能。上述用户触达、用户启动、注册成功都能表明用户希望或者正在尝试你的产品，若用户没有体验产品的核心功能，就没办法给客户带来实际价值，因为解决不了用户的任何问题，用户也没有回来的动力。

激活阶段的目的是实现用户转化，激活是用户留存的前提。因此，想要提升用户留存，首先要考虑这些流失的用户是否真的体验了产品的核心功能，也就是说是否被真正激活了。只有被激活的用户多，才可能有更高的留存率。

如何衡量用户激活

不同企业对“激活”的定义不同，我们将新用户完成核心功能的初步体验称为“激活”。一般而言，互联网金融行业用户绑定一张银行卡，电商用户完成一次购物，视频平台新用户完成了 3 个视频的播放，图片处理产品的用户要进行若

干次的图片编辑及发布，才被视为“激活”。

不同的产品定义“激活”有着不同标准，甚至在产品发展的不同阶段也有一定差异，但归根到底，一定要让用户体验到核心功能。

不容忽视的激活因素

激活用户，以下三个因素不容忽视。

1．赋予用户 Aha! Moment。

Aha! Moment 是新用户发现产品内在价值的时刻，这一时刻让新用户成为黏性用户，新用户就有很大可能会被留存。如果用户体验产品时没有体会到价值，就很难留下来。就像对一个视频网站来说，用户只是到首页逛了一圈，并没有真正观看过视频，是不会有 Aha! Moment 的。

显然 Aha! Moment 是影响用户留存的关键点。实现此时刻并非滔滔不绝传递“我们有上千位企业客户”、“我们这里有最全面的影视资源”等信息，这些话也许会让用户对你的产品感兴趣，相信你的产品具有一定的价值，然而，真正的 Aha! Moment 一定是用户亲身体验的结果。

在产品流程设计中，应该尽量让客户以最低的代价体验到 Aha! Moment，否则夜长梦多，容易流失。

2．找到你的 Magic Number（魔法数字）。

所谓 Magic Number，就是用户在执行了某些操作序列后，更容易成为一名忠实用户。企业为赋予用户 Aha! Moment，通过大量工作来找到企业发展的 Magic Number，并将其作为努力的目标。Magic Number 是企业的“可行动指标”，是用户了解产品的价值并转化为终身用户的“奇妙界点”，这个界点能让用户感受到产品的价值。

“10 天内交到 7 个朋友”是 Facebook 用户数量从零增长到十亿过程中的卖点，让用户完成该动作就成了 Facebook 内部全体员工的核心目标，正是因为对这个目标的坚持，他们最终汇聚了 10 亿用户。

Twitter 用户流失率曾达到 75%，时任增长团队的产品负责人约什•埃尔曼（Josh Elman）深入地研究余下 25% 用户留存的原因，发现这部分用户群体关注的人都超过 30 个。于是，Twitter 将“关注 30 人”作为 Magic Number，注重引导新用户

关注好友，并为其推荐好友关注，以此提升用户留存率。

当企业确定 Magic Number 后，可以采用引导、推荐等方式。做产品的朋友非常喜欢聊“如何像 Facebook 一样，找到属于自己的 Magic Number”这类话题，实际上 Magic Number 不是在灵光一闪时出现的，需要通过大量的数据分析来发掘。这里要提醒的是，Magic Number 只是一种表象，背后的原因是用户只有在进行了这些操作后，才真正体验了产品的价值。如果 Facebook 的新用户没有好友，就没有人互动，显然没有趣味性。如果 Twitter 用户没有 follow 足够多的人，就不能看到足够的更新，用户也就没有刷 Twitter 的动力。

3．把不同的环节通过 ID 串联起来。

用户激活的过程是一系列操作，这时需要考虑把用户的行为通过 ID 串联起来，以方便优化分析。比如，有些产品只记录了用户注册后的行为，与注册之前的行为不对应，就可以通过匿名 ID 和注册后 ID 关联的方式，来实现数据打通。在有些线上线下结合的场景，有时候甚至需要通过人工的方式实现 ID 串联。用户激活的过程一般可以通过一个漏斗表现出来，具体过程为：访问首页 → 浏览商品 → 注册 → 下单。

如何提升用户激活率

上文提到，激活用户的核心在于让用户尽早尝到产品的“甜头”。一切运营活动产生效果的前提是有一个好的产品，当有了好的产品后，如何提升用户激活？不妨尝试以下途径。

1．减少干扰。避免繁文缛节，让用户尽快使用核心功能。有些产品要求用户填写大量表单，或者首页做得过于复杂，让用户望而却步，很容易选择离开。运营人员应该考虑到如何让用户先用起来，以后再去补充必要的信息。

2．提升性能。如果用户打开产品出现卡顿，就不要指望用户的高激活率。

3．增加引导。当用户接触产品后，要引导用户到最关键的操作步骤上来。

4．人工接入。对于 2B 服务来说，用户注册试用后，即显示出了强烈的需求信号。然而产品复杂性需要有专门的咨询顾问对客户进行解答，只有不断为客户解疑释惑，才能让用户真正体验到产品价值。

以上只是简单列举几个可行的方式，企业还应具体问题具体分析。

案例：使用漏斗分析和用户分群来激活用户

以神策数据官网为例，官网访客往往经过访问首页→点击“申请试用”→提交“申请试用”→登录 Demo 并进行体验操作等环节。我们认为只有走到第四步（试用产品）之后才算是“激活用户”，因为用户只有亲自试用后才知道神策数据是一款什么样的产品。

我们可以通过漏斗分析模型记录用户路径中详细的转化过程。神策分析刚上线时，从提交“申请试用”到体验 Demo 的比例只有 41%，如图 4-2 所示，经过优化和人工咨询服务跟进之后，这个比例提升到 2015 年年初的 70% 以上，现在的比例更高，如图 4-3 所示。

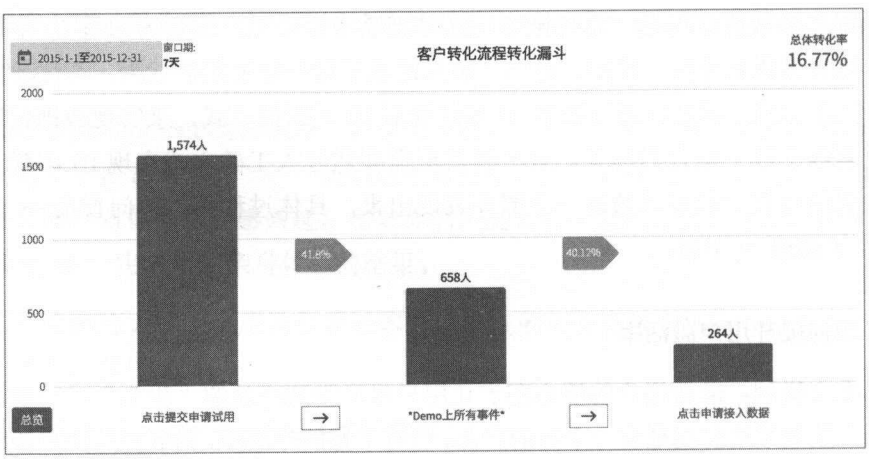


图 4-2　神策分析产品刚上线时，官网访客的转化情况

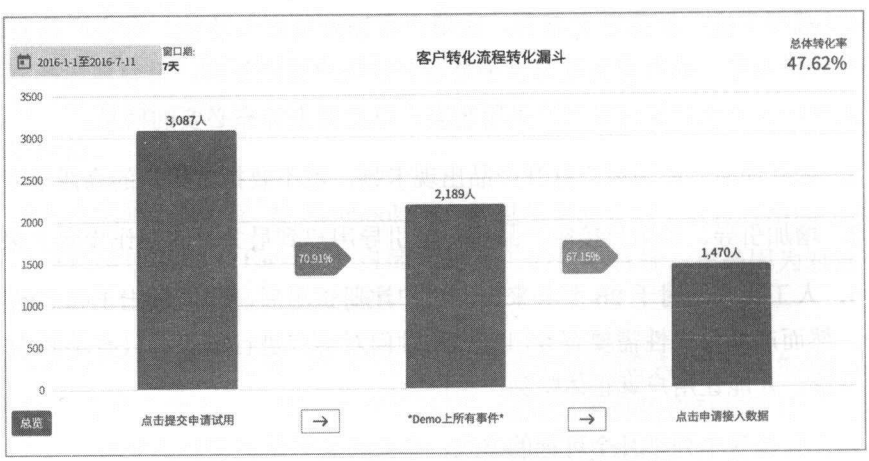


图 4-3　经过优化后，官网的用户转化情况

我们还可以通过“用户分群”数据分析模型，将满足激活条件的用户和不满足激活条件的用户区分开来，并跟踪他们的差异，找到合适的 Magic Number，如图 4-4 所示。

图 4-4　将符合激活条件的用户进行筛选并分群

留存（Retention）

留存用户量和留存率体现了应用的质量和保留用户的能力。复购率、次日留存、周留存都是与留存相关的指标，企业根据产品和业务特点定制统计指标，以监控用户留存或流失情况，做到关注留存、提升留存和利用留存。

关注留存：客户获取成本（CAC）小于客户生命周期价值（LTV）

LTV 和 CAC 是商业模式可行性中不可或缺的两个变量。业界普遍认为，CAC 要明显小于 LTV 才能算平衡的商业模式，在企业服务领域更甚，LTV 大于 3 倍的 CAC 才算得上企业良性发展，这已经被认为是全球通用的固定法则。

中国电子商务中心提供的公开数据显示，商家获得新用户的成本是维护老用户的 5 ～ 10 倍。居高不下的获客成本，对客户留存率提出了更高的要求。因此，企业要关注留存。用户培育应该像滚雪球一样越滚越大。像猴子掰玉米一样，一面努力开拓新用户，一面有大量用户流失，显然是企业培育客户的大忌。此外，

我们要防患未然，要有预警流失用户的意识。企业运营人员可以尝试找到流失用户的一些相同行为特征，借助第三方数据分析平台，分析用户在流失前是否有类似的行为，如体验了烦琐的售后流程、卡顿的视频等；或者发现一些奇妙的数字，如用户多少天不回访一般都会流失……

发现这些共同点后，我们可以在运营过程中重点对糟糕行为体验进行优化，或适时采取召回策略，保证用户多少天内出现第二次回访。

提升留存：降低流失率

正因为获取新用户困难，所以我们要想办法让用户尽可能留下。留存率应该提升到什么程度呢？不同的产品类型，并不具备可比性。比如对于微信来说，一个用户每个月有 25 天活跃，可能就说明留存度不够。而对于京东来说，一个用户每个月有 4 天活跃，就是一个很好的用户。我们都想让留存率涨到不能再涨，下面将分三点介绍如何提升留存。

1．精准信息推送

一条关于降价的 APP 推送可能会“刺激”用户购买存放在购物车里很久的商品，一条“收益率高达 9.50%”的短信推送可能会让客户立刻进行二次投资。

信息推送首先要“精准”，要精准聚焦在业务转化率薄弱的环节。企业运营人员应事先根据用户行为进行用户画像——勾勒某用户群体在特定研究范围内的行为全貌，将具有类似属性与行为特点的群体“圈”出来，再对目标人群进行精准的信息推送，以刺激用户留存。

其次要合理把控推送频率及内容，避免骚扰用户。要合理安排推送时间，并根据产品使用的频次决定消息推送的频率，避免向用户推送不适合或者无效的信息，以及要尽可能推送用户感兴趣的内容。

2．运用 Magic Number

前面介绍的 Magic Number，其实就是用来让用户在完成某一系列操作后，真正体验到产品的价值，更容易留存下来。我们提升留存的思路，不要局限在如何让已有的用户更活跃，而是让新来的用户尽快地进入最佳的体验状态。

3．流失用户挽回

用户流失原因有很多种，我们可以有针对性地挽回用户。挽回用户，即运营

人员告诉用户“我的产品能够满足你的需求”。告知用户的渠道很多，其中精准推送是挽回流失用户的策略之一，APP 推送、社交媒体推送、短信、邮件都是可以触达用户的沟通渠道。我们可以通过用户分群，根据用户属性及用户历史访问情况，将部分用户圈出，然后有针对性地进行挽回操作。比如针对两个月前有购买行为，最近一个月没有到访的用户群，特意给他们推出折扣优惠。

利用留存：读懂用户留存，最大限度延长用户生命周期

企业要尝试读懂你的用户留存，可通过日留存率、周留存率、月留存率等指标监控应用的用户流失情况，关注用户留存变化与留存规律，并在用户流失前提前采取相应措施，激励这些用户继续使用。如此无限延长用户的生命周期，最大化用户生命周期内的价值。

案例：Worktile 利用留存规律，销售业绩提升 3% ～ 5%

Worktile 是一个团队协同办公工具，致力全面打造互联网时代的一站式企业协同管理平台。企业通过数据分析发现两个有关留存的规律。

1．企业发现 70% 的用户会在一天内完成从“注册页面完成”到“成功创建团队”的操作，而超过一天的流失率会大大增加。

2．14 天的免费试用期中第 7 天还在试用的客户通常是高潜客户。

发现了这两个规律后，Worktile 团队做出针对性策略，即在用户注册的第二天，销售会进行电话拜访，以增加用户黏性，并对高潜客户进行电话追踪，提高成单效率。通过优化，他们的销售业绩提升了 3% ～ 5%。

引荐（Referral）

关于引荐，最早的案例要从 1996 年 Hotmail 刚出现时说起，作为最早的 Web E-mail 创新者，如何才能吸引更多的新用户去使用？ Hotmail 的投资人想到了一个主意，在签名档里加入一个链接，询问用户有没有兴趣获取一个免费的 E-mail，用户点击链接后，就会跳转到注册页面。通过这种方式，Hotmail 在短短一年半的时间里，增长到了 850 万个用户，并于 1997 年 12 月以 4 亿美元的价格卖给了微软，成为互联网历史上的传奇。

随着增长黑客的兴起，每个企业都希望自己的产品能够实现病毒式传播，让

产品发挥出 10 倍、20 倍甚至 100 倍的威力。有些产品天然具有病毒性，用户会基于自身诉求而进行自发传播，例如 PayPal。

PayPal 团队把用户增长作为第一要务，他们相信网络效应。产品的影响力是用户数的平方，如果你的用户数是对手的两倍，那么你的产品影响力就是对方的 4 倍。具体怎么去吸引用户呢？正如 PayPal 的病毒式传播，当一个 PayPal 用户向朋友转钱，他会促使朋友也注册 PayPal。PayPal 还通过人工的方式促进这一传播速度，比如 PayPal 用户每引荐一位朋友注册，就可以得到 10 美元的奖励，最终 PayPal 实现用户爆发式增长。

并非所有的产品都适合病毒营销，要回归价值

病毒营销对于一款产品的增长很重要，但绝大多数产品并不具有网络性质。无论是企业产品、市场，还是运营工作，最终都要回归产品的本质。即使是病毒式营销成功的 PayPal，用户进行推荐的前提也是产品要足够好。否则用户没有推荐的动力，他们不愿意为赚钱去推荐一款糟糕的产品而被朋友看不起，或者他们在领取 10 美元奖励后毫不犹豫地离开。这个推荐奖励对 PayPal 的增长至关重要，也正是因为产品的价值，商家愿意把 PayPal 的标识链接放在商品详情里。

总之，一个好的产品会说话，最直接的就是口碑力量。要让用户说话，把权利交给用户。

留存率不代表用户忠诚度高

前面提到了“留存”，留存与利润相关，但留存与新用户增长率直接关联度较弱，它不能真正衡量出用户的忠实行为。用户可能是因为昂贵的转移成本，不得已放弃其他产品，而被迫停留在你家产品上。

例如，企业对正在使用的云服务提供商颇有微辞，当用户想弃用该提供商的服务时，发现迁移数据并不轻松，在不同服务商之间的数据迁移成本甚至高过迁入成本。为保证业务不受影响，用户最终不得已维持原有云服务商的合作关系。

与留存相比，NPS 更能直截了当地判断用户对企业的忠诚度。盲目追求大量用户，NPS 不理想也会造成用户流失率高居不下。总之，NPS 已成为企业必备的衡量标准。

关注 NPS：用户会主动推荐你的产品吗

中国电子商务中心提供的公开数据显示，一个满意的用户会带来 8 笔潜在生意，不满意的用户可能会影响 25 个人的购买意愿。

在贝恩咨询公司的弗雷德·赖克哈尔德（Fred Reichheld）所著的《终极问题 2.0：客户驱动的企业未来》一书中指出，用户忠诚度是区分“良性利润”与“不良利润”的指标。不良利润是以恶化顾客关系为代价赚取的利润，这种商业模式会造成用户流失并转向竞品，甚至会阻止身边其他人使用；而“良性利润”真正践行“以用户为中心”，会提升企业口碑，促成用户向周围人推荐，促进产品可持续增长。他总结道，NPS 已从一个指标提炼上升到一个系统。

因此，你不妨问客户一个问题：是否愿意将产品推荐给你的朋友或同事？当你感受到 Google 良好的搜索体验时，你自然会向周围人推荐 Google。当更多用户自发去推荐企业的产品或服务时，就会自动提升企业良性利润占比，提高口碑和销售额，从而促进企业增长。

我们一直很关注 NPS，在我们 400 多家付费客户中，哪些企业只是自己在用？哪些企业会为神策数据推荐新客户？哪些客户在私下或公开对神策数据进行负面评论？这些我们心里有谱。我们把愿意做推荐的客户数量减去有过负面评论的客户数量，得到的就是 NPS。一个好的产品，NPS 应该在 50 以上。只有实现从产品自传播到再次获取新用户，企业运营才能形成一个螺旋式上升的轨道。

营收（Revenue）

任何一家企业的最终目的都是为了盈利，这一点毋庸置疑。烧钱可能会维持公司发展一段时间，但最终要实现的是正增长。因此，企业十分关注付费转化率、销售额、平均客单价、新增付费用户数等指标。

依旧以 PayPal 为例，看它是如何做的。PayPal 在刚推出时承诺不向用户收费，因为若强制把用户变为付费用户，则极有可能造成用户流失。他们在数据分析时发现，有 10 万个账户每半年接收信用卡付款超过 500 美元，这些账户的交易额占总交易额的 75%。于是他们就采用了一个策略，如果用户半年内的信用卡收款超过 500 美元，超出的部分就进入“待定”状态，卖家要么选择升级付费账户，要么系统就会把钱退还给买家。

该策略自然引发用户抱怨声一片，但当用户权衡利弊后，不到一个月时间，95% 的目标用户都完成了升级。由于用户使用信用卡付款时，PayPal 要给信用卡公司 2.5% 的手续费，而借记卡只要支付不到 1% 的手续费。为此，他们限制了用户每年的信用卡支付限额，促使用户使用借记卡进行付款。这两步做好之后，PayPal 成了一个赚钱的生意。

我在知乎《PayPal 早期是如何通过 Growth Hacking 成长为独角兽的》系列文章中，详细介绍了 PayPal 如何向高交易量卖家收费的非常成功的转型故事，有兴趣的朋友可以自行查阅。

案例：金融平台打通行为数据与业务数据，提升产品营收

M 公司是国内领先的金融垂直搜索平台，处于贷款机构和用户的中间环节。营收一直是公司业务层面的数据。M 公司本身不提供产品，最终目标转化会跳到第三方平台，为实现端到端的分析，线上行为数据和最终业务数据需要对接与贯通。

例如，对于市场部来说，要以最终的收益来考核 ROI。用户来源于不同渠道，企业需要系统地分析哪个渠道带来的用户最为优质。

M 公司打通了线上数据和线下数据，实现三方业务数据导入和 BI 系统集成。市场部可以获取从推广到最终业务转化的完整数据，准确高效地计算投入产出比，衡量不同渠道的投放。最终市场部实现精准拓展与精细化运营，不断提高产品营收和商业价值。

数据驱动产品改进和体验优化

页面底色设置是绿色好，还是蓝色好？信息流应该放置在左侧还是右侧？购物车容量应该设置多大？面对抉择，网站设计师给了 100 多种选择绿色做背景色的理由，Facebook 将信息流放在了左侧，该听取和效仿吗？

产品生命周期经历了诞生、成长到成熟，最终走向衰亡的过程。企业为延长此过程，需要对产品不断地进行优化和改进。前面数次提到，“拍脑袋”做决策依赖的是过往经验，可能“拍”对，也可能“拍”错，“因果驱动”可能会延误最佳决策时机，产品改进方向千千万万，而数据比一切都可靠。

MVP[1] 理念强调以最小的代价推出一款最小可用的产品，然后在此基础上进行迭代。在这个迭代过程中，数据分析是非常关键的。我们不应该陷入“拍脑袋”→“研发功能”的循环中，而是要引入数据分析环节。

当一项新的产品功能上线后，后续工作并不是一股脑地继续开发新功能，而是要阶段性地停下来用数据评估新功能是否达到了预期效果？通过数据找到薄弱环节或者不合理的环节，持续将产品正向迭代，如图 4-5 所示。

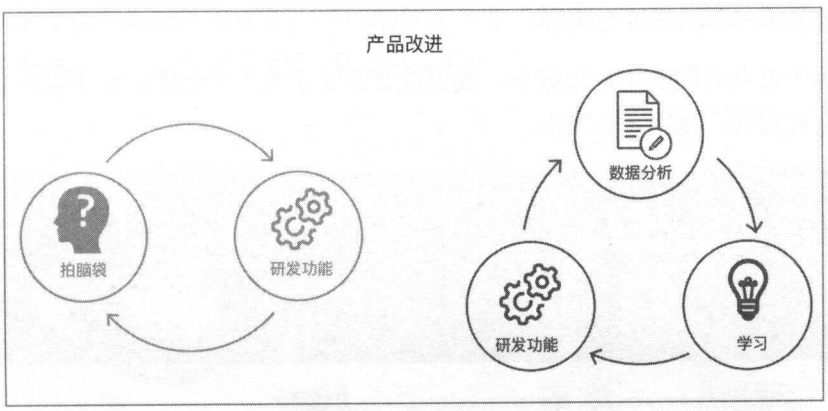


图 4-5　数据分析驱动的迭代

案例 1：产品改版——跟风借鉴 VS. A/B 测试

Facebook 的 Growth Hacker 一直专注用产品和运营数据来验证结果。其中一次改版将左侧繁杂的导航栏变得简洁，Feed Story 里面的图片被加强突出。之所以这样做是因为产品经理认为图片是整个 News Feed 中最为吸引人的内容。新版的 Facebook 页面焕然一新，国内许多知名社交网络，如微博等开始纷纷效仿。

一切都看似合理，然而，这支 30 人的精英团队接近一年的工作在数据面前低了头：DAU、用户参与度、在线时长都在持续下滑。最终 Facebook 改回了之前的旧版本。

Facebook 任何产品优化，哪怕只是在局部小规模的优化，都要进行灰度发布和数据验证，产品人员一直密切关注任何有关用户体验的数据。显然，国内许多企业缺乏数据驱动的意识，此次“借鉴”事件被媒体批为“自掘坟墓”。

相比之下，36 氪则很好地践行了数据驱动的理念。在技术产品中，A/B 测试

[1] Minimum Viable Product，最小化可行产品。

是最常用的数据驱动实践之一，将产品的用户流量分成两组属性相似的用户群组，在同一时间维度内有两组群体在使用产品，通过收集各群组的用户体验数据和业务数据，最后分析评估出最好的产品版本。

案例 2：神策数据官网改版

以我们的一次改版为例，为了更好地评估改版效果，我们进行了 A/B 测试。让 50% 的访客进入旧官网，让 50% 的访客进入新官网。然后我们对比从访问首页到体验 Demo 的漏斗转化比例。

由于首页访问的用户量较少，我们就多积累了几天的数据，结果发现总体转化并没有提升，如图 4-6 所示。

图 4-6　神策数据新旧官网对比图

图左侧是旧官网，图右侧是新官网。通过漏斗分步骤发现：用户浏览到提交试用申请的比例明显提升了，但申请后登录 Demo 的比例却下降了。经过排查，我们发现新官网申请完成的跳转链接做了错误配置，跳转到了旧官网对应页面，导致用户体验的落差。修复问题后，新官网比旧官网的转化率提升了 50% 以上。如果没有做精细化的数据分析，就很难发现问题。

数据驱动商业决策

大数据在企业的商业决策和商业价值决策中扮演着重要角色，这已经是普罗

大众的共识。数据驱动是企业发展最基本的要求，百度也非常重视“用数据说话”，我们来看一个曾经发生在百度知道的例子。

2007 年我加入百度知道，当时的产品经理告诉我，每年暑假，百度知道的用户量和提问量都增长很快，因为一些学生需要完成假期作业，有问题就上百度知道提问，同学之间传播得也很快。2012 年，百度知道的团队成员对用户的提问类型做了分析，发现学生提问占了 10%。

于是在百度知道的 APP 上，专门增加了一个“问作业”的菜单。结果，作业相关的提问量一下上升到了 30%，如图 4-7 所示。

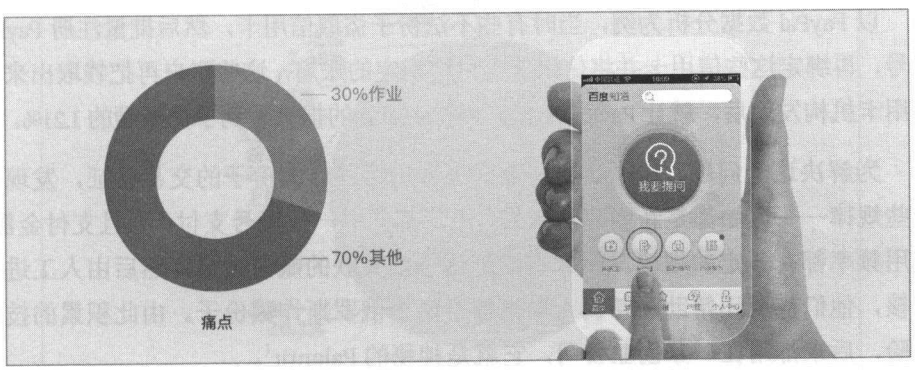


图 4-7　作业帮产品

于是团队开始考虑开发一个独立的 APP，用于学生提问。

终于，一款叫“作业帮”的产品诞生了。产品诞生后的半年时间，与作业相关的提问量提升了 4 倍。这就是一个典型的数据驱动商业决策的例子。

在硅谷，用数据驱动商业决策的案例，更比比皆是。

案例 1：Google 为员工提供免费午餐背后的数据意义

Google 强大创新力的背后是数据。《重新定义公司：谷歌是如何运营的》（*How Google Works*）一书在“决策：共识的真正含义”一章中指出，制定决策的方式、时机和实施决策的具体方法与决策本身同样重要，数据越翔实，信息就越清晰，推理就越高效。作者指出要用数据做决策，观点的背后一定是数据的支撑，观点不能在缺乏数据支撑的情况下单独存在。

Google 为员工提供免费午餐这一福利背后，也有着缜密的成本计算。Google 为此需要支付的成本不足 10 美元，而员工为此却减少了外出就餐的时间浪费，对于

时薪较高的员工，更是有效延长了工作时长，降低了生产力的浪费；而且不少员工就餐时会邀请朋友，还能兼顾招聘，成本核算后发现比猎头挖人成本低很多。类似的，谷歌用一套非常严谨的以数据为驱动力的决策流程，优化企业治理和商业决策。

案例 2：PayPal 的反诈骗之战

时任 PayPal 的 CEO 彼得·蒂尔（Peter Thiel）格外看重数据的价值，这与他创办过 Thiel 资产管理公司的经历有关。《支付战争》一书提到，当员工向彼得·蒂尔汇报工作时，若其汇报内容有丰富的数据做支撑，彼得·蒂尔会赋予他更大的自由权。

以 PayPal 数据分析为例，当时有些不法份子盗取信用卡，然后批量注册 PayPal 账号，再绑定这些信用卡并将信用卡支付给指定的账户，这些账户再把钱取出来。信用卡机构发现后，就让 PayPal 赔偿。PayPal 因此的损失达到了交易额的 1.21%。

为解决这个问题，PayPal 通过数据去分析这些诈骗份子的交易特征，发现有一些规律——比如都是新创建的账号，用户都向同一批账号支付，并且支付金额、使用频率都有一定的规律。因此它们把这一批关联的账号锁定，然后由人工进行审核，他们甚至还帮助联邦调查局抓捕了两个俄罗斯诈骗份子。由此积累的技术经验，后来都用在一家创业公司，它就是神秘的 Palantir[1]。

为了阻止程序自动注册账号，技术人员想到了一个办法：他们生成一些图片，图片上印有扭曲的字母，并且有一些背景划痕，用户可以识别这些字母，但是机器很难识别，这样自动注册就行不通了。现在这种图像识别码的方式已经非常普及，甚至到了滥用的地步，而它最早就是 PayPal 发明的，至少是最早商用的。

通过这两个措施，诈骗损失从交易额的 1.21%，降到了 0.48%，导致诈骗份子们只能“跑”到竞品上作案。

数据驱动落地企业，要从管理者做起

实践是检验真理的唯一标准，数据是验证实践的科学依据。数据驱动是最先进的生产力，让数据驱动落地企业，最为有效的方式是从上而下地推动。管理者推动的前提是其自身具备数据意识，能够认识到大数据的意义与价值，意识到数据管理不善可能带来的危害。若某下属费劲整理了一批数据报表，领导却对此无

[1] Palantir，神秘的硅谷数据分析公司。一些外媒称该创业公司帮助美军捕杀了奥萨马·本·拉登（Osama bin Laden）。

视，仍坚持拍脑袋做决定，这样企业数据发展就会陷入绝境。

管理者应主张建立企业的数据决策文化，在整个企业层面建立一种以客观的数据为决策依据和衡量标准的价值观和制度体系，普及数据意识，并建立相应的组织架构。

作为一个初创企业，神策数据的一切决策也是基于数据。神策数据的官网访问量较小，一些简单的数据分析也能为商业决策指明方向，如图4-8、图4-9所示。

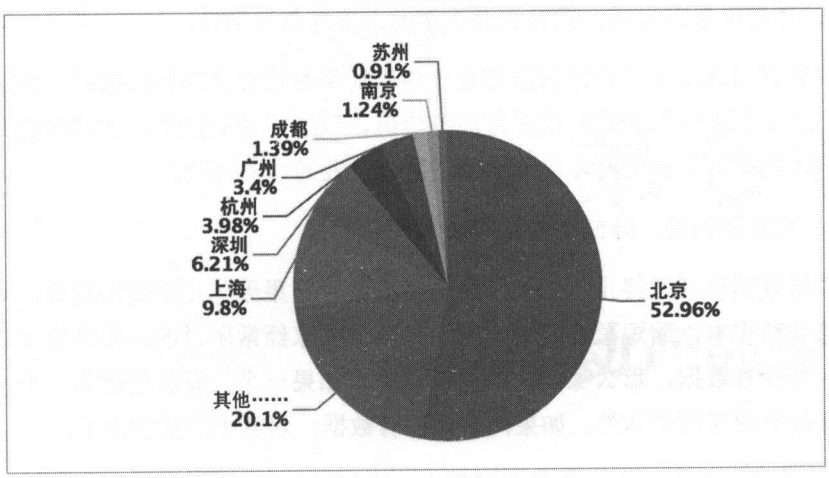


图4-8　2015年12月，神策数据官网访客来源

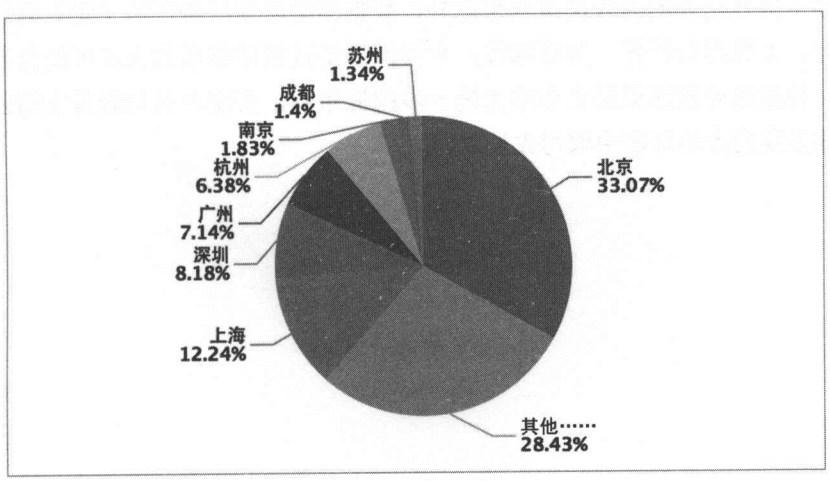


图4-9　2016年6月，神策数据官网访客来源

通过神策数据官网不同时期的用户分布情况，我们能够发现一些特点：之前

北京占据了客户来源的50%以上，近来北京所占份额下降，各地域呈现均衡状态，说明市场宣传已经触达更多区域，同时官网访客的情况也辅助了我们对分公司选址等的决策。

数据驱动商业决策的价值

综上所述，我认为数据驱动商业决策实现了以下三大价值。

1．透过现象看本质，提升企业决策的准确性与科学性。

大数据时代改变了曾经仅依靠企业内部业务数据优化决策的情况，通过洞察“大”“全”“细”“时”数据背后的价值，赋予企业更加全面和准确的商业洞察力，大幅提升企业的商业决策水平，降低企业经营的风险。

2．用数据说话，降低企业沟通成本。

“用数据说话”终止了企业团队之间因某一结果无休止争论的局面，降低了团队的沟通成本，缩短了企业研发时间。就像大家经常所说的，在争论中，如果两个人都没有数据，那么嗓门大的通常获胜。如果一个人有数据而另一个没有，那么有数据的通常会获胜。如果两个人都有数据，那就不需要争论了。

3．赋予企业全面准确的商业洞察力，实现智能商业预测。

除了提升企业的商业决策水平之外，数据驱动商业决策还赋予企业商业预测的能力。大数据分析像一架望远镜，企业通过望远镜能够看到未来可能会发生的情况。智能商业预测帮助企业率先进一步挖掘市场，实现产品与服务上的创新，在诡谲多变的市场环境中取得先发优势。

第5章

数据驱动产品智能

我们在第 4 章讲解了如何通过数据来驱动产品和运营决策，这可以说是在互联网产品中比较广泛的一类数据驱动场景。正如在第 2 章所提到的，在我看来，数据驱动决策只能发挥数据 20% 的价值，甚至更少，而数据更大的一个价值在于驱动产品智能。

所谓智能，我把它归结为这么一种模式：首先我们要有数据，然后在数据上套用某种算法模型，最后再将结果数据反馈到产品中，这样的产品就具备了一种“学习”能力，这就是我说的产品智能。数据驱动产品智能模式如图 5-1 所示。

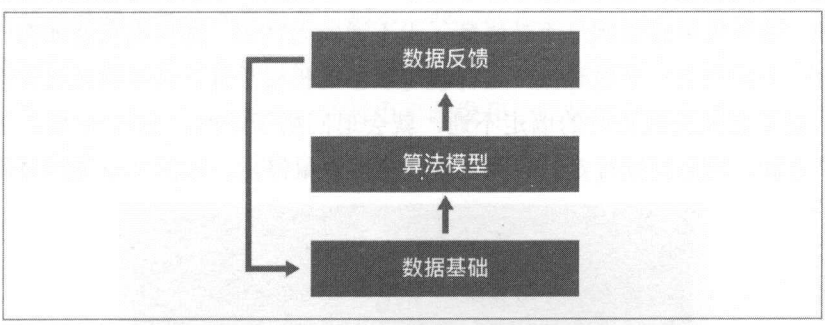


图 5-1　数据驱动产品智能模式

数据驱动决策将分析结果用于人的决策使用，而数据驱动产品智能更加强调数据的处理结果是给机器用的，并且这种数据分析的算法往往更加复杂，本身具有可以自我迭代的特点。不管是百度搜索引擎根据用户的点击情况自动调整排序，还是今日头条根据你看过的新闻给你推荐相关新闻，都属于这种情况。

关于数据与智能的关系，有一个有趣的故事。故事的主人翁是克劳德·香农，信息的单位比特——bit 就是他命名的，他是信息论之父，我们所有的信息技术，

都是建立在他提出的理论基础之上。

克劳德·香农的机械鼠——“忒休斯”实验能够给我们更多启发。

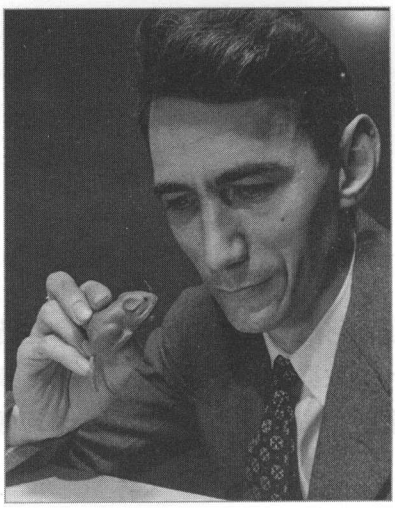


图 5-2　克劳德·香农和他的机械鼠“忒休斯”（图片来源于互联网）

1952 年，克劳德·香农在一次会议上展示了他制造的一只老鼠——一只可以走迷宫的机械鼠。这只老鼠有三个轮子、一根磁铁，以及铜线做成的胡须。通过胡须，老鼠可以感知到是不是碰到了走不通的迷宫墙。迷宫地板背面有一个机械手臂，上面也有一个电磁铁，这样就可以移动机械手臂，带动机械鼠在迷宫里走动。如果老鼠发现正对的墙走不通，就会退回格子中间，旋转 90 度，去尝试下一个方向，然后继续行走。直到走到终点，老鼠停止，如图 5-3，图 5-4 所示。

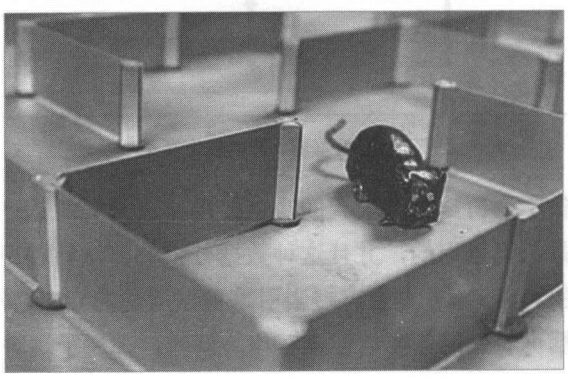


图 5-3　“忒休斯”在迷宫中走动（图片来源于互联网）

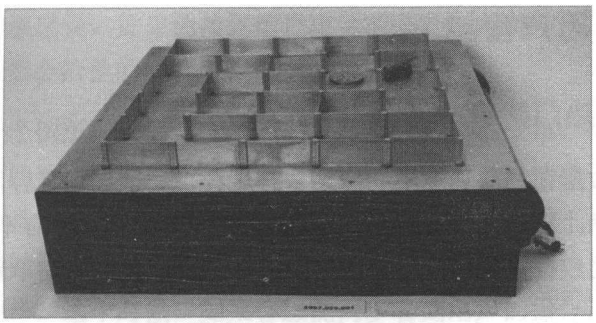


图 5-4　“忒休斯”走到终点（图片来源于互联网）

神奇的是，如果把老鼠重新放回到起点，它会直接沿着正确的路走到终点。如果调整了中间的线路隔板情况，老鼠还会重新探索路线，正确走到终点。这只老鼠是怎么做到智能的呢？

在整个电路中，香农用 50 个继电器控制机械手臂的移动，又用 75 个继电器来记录老鼠探索的每面墙是否能走通。这只老鼠显然是学习了迷宫路径，才能重复正确的路径。它通过继电器记录了路径状态，也就是说，老鼠通过掌握了更多的数据，从而实现了这种智能。这里甚至没有牵涉对数据的处理，仅仅是记忆这些数据，就可以拥有智能了。

这应该是最早期的产品智能，如今一些智能类的产品应用更为复杂。为了实现产品智能，我们需要完善的数据平台及针对场景的算法模型。接下来我们重点讲解数据平台和机器学习算法，并对用户画像和个性化推荐两个典型的产品智能场景进行描述。

图 5-5　香农掀开迷宫的底板，展示机械手臂与电路设计（图片来源于互联网）

数据平台及用户智能

如何计算热门榜单

用户行为数据在产品功能中的应用多种多样，一个典型且容易理解的例子是各类榜单的计算。几乎所有的小说、视频、音乐等内容网站都有不止一处的榜单，这些榜单主要的数据依据就是用户的行为数据，下面我们来简单看看这一过程是如何进行的。

首先，我们需要采集到用户在产品上的各类行为，例如搜索、浏览、播放等行为，这些行为需要通过 APP 或者浏览器发送，然后到达数据接收服务。

紧接着，我们需要对这些数据进行清洗。行为数据中会存在大量的非法数据，包括机器访问（例如搜索引擎爬虫）、非正常用户访问（例如靠刷量产生的用户），或者干脆直接就是程序模拟的行为数据。这些数据会导致榜单数据不准确，因此需要在这个阶段进行清洗。

由于要兼顾榜单的时效性，实时的数据清洗一般只能利用一个较短窗口期内的数据来做决策，并且无法回溯数据。例如对于一个特定 IP 的访问，可能处理了 500 条之后才能判断来自该 IP 的访问是非法的，但是这个 IP 的行为可能已经被用于之前榜单的计算了。

在经过这一阶段之后，我们就可以拿行为数据来计算实时的热门榜单并将其更新到产品上。根据产品需求的不同，可能是秒级的实时更新，也可能是 5 分钟甚至半小时级别更新。

实时的行为数据不能在计算完成之后就丢掉，而需要被持久地存储。因为除了实时的热门榜单，一般的内容网站往往还会提供周榜、月榜等周期的榜单。这些榜单需要更长周期的数据以及更复杂的策略，例如综合考虑播放量、播放时长、评分等信息。并且，在这一阶段我们有了更丰富的信息，可以对数据进行进一步的清洗，例如可以找出那些长期进行刷量的黑名单，以进一步提高数据的可靠性。由于更新周期足够长，在最终的结果被使用之前还可以加上人工的编辑审核，以确保榜单结果符合产品运营的需求。

客服系统中的行为数据

用户行为数据和客服系统的结合也是一个非常实用的例子。例如对于一个在

线订票网站的客服来说，如果他接到用户电话的那一刻，就知道这个用户近期做了什么操作，显然会给他的客服工作提供很大的帮助。

类似于第一个例子，我们同样也需要进行行为数据的采集、传输、存储等。不过，这时复杂的数据清洗工作不是必要的，因为绝大多数情况下的非法数据都不会针对某一个具体的人，而少量的脏数据并不会对应用造成干扰。这时我们更希望尽快让数据能够被客服系统查询到，因为客户有可能会在他打电话的同时也进行一系列的操作，这个时候，如果客服能同步看到用户的操作，显然会更有帮助。相反，如果因为不必要的数据清洗导致数据延迟了几分钟，则完全是对资源的浪费。

另外，在这个应用场景中，我们并不是产出一份固定的数据，而是需要实时地根据当前客户的 ID 来查询客户近期行为，这一需求对于数据存储系统也会有一定的要求，即存储系统需要具备根据客户 ID 来进行高并发查询的能力。

值得注意的是，用户在客服系统中也可能会产生一系列的行为，例如客服协助用户进行了退款操作，实际上即是用户发生了一次退款行为。显然，这些行为也会和其他用户行为一样被真实地采集、传输、存储下来，当下一次用户打客服电话时，客服能参考这些行为。

为什么需要数据平台

除了上面提到的两个例子，还有无数个或简单或复杂的数据应用，例如个性化推荐、风控、精准营销等。即使是热门榜单这个看起来很简单的功能，也会随着产品的迭代不断产生新的需求，例如某一个版本可能需要不同打分依据的榜单，或者是产品经理在某一天希望针对不同板块来设计不同榜单的策略。很显然，在功能实现上，我们不可能对于每一个数据应用都单独进行数据的采集、传输或者存储。

虽然产品功能是易变且难以预测的，但是所有这些基于数据的功能依然会有很多的共同点，尤其是它们对于底层基础数据的处理和访问的需求，基本上是稳定不变的。针对这一部分的需求，我们可以抽象出一个比较通用的数据平台，以减少不同数据应用中的重复数据处理过程，让不同的数据应用更专注于业务相关需求的实现，不用纠结于数据从哪来、数据如何访问等重复枯燥的问题。

相对于上层数据应用的易变性，数据平台更要做到以不变应万变，用一个比

较通用的架构适应众多的数据需求。需要注意的是，不同的数据应用对于数据的需求并不是完全一致的。例如，我们上文提到的两个例子虽然都是行为数据的应用，但是很显然它们在数据的时效性、准确性、访问能力上都有着完全不同的需求。因此，设计一个灵活而通用的数据平台架构是相当不容易的，需要平台的设计人员对技术架构、产品业务都有非常深入的理解能力以及抽象能力。

数据平台提供的能力

为了支撑多种多样的数据应用，一个数据平台需要具备对数据的灵活处理能力，包括接收、清洗、存储、计算、查询等，整个过程中既要足够高效，又要考虑通用性。

数据接收

典型的数据平台架构如图 5-6 所示。用户行为数据是一个天然的实时数据流，因此在数据接收层面，必须要提供可靠、实时接收的能力。无论采集端以什么方式进行，最终都应当使用统一的接口将数据发送到数据接收层。

接收层通常不需要复杂的处理逻辑，更重要的是保证接收服务的可靠性以及可扩展性，在面对突发性的流量增长时不会造成数据丢失。

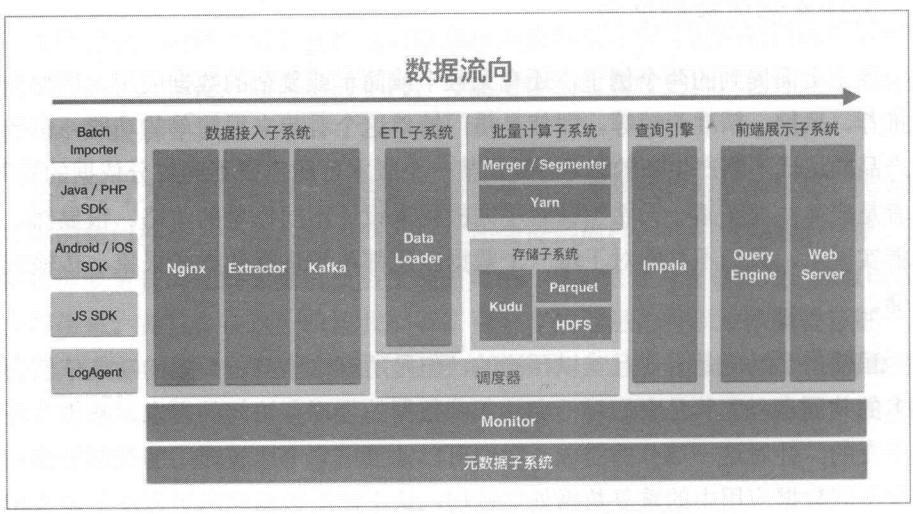


图 5-6　一个典型的数据平台架构

实时订阅

数据到达接收层之后，会被立即写入一个消息队列中，以对下游提供订阅服务。这里基本上没有任何额外的处理逻辑，因此延时是非常低的，通常都在 1 秒以内，甚至只有几十毫秒。

举个例子，一个用户在产品首页搜索了一个关键词“鲜花”，立即会有一条代表此行为的数据被发送到接收层。而后续的一个针对用户近期搜索行为提供推荐服务的模块则可以在 1 秒钟内拿到这个行为的数据，并且在用户访问下一个页面的时候及时提供“鲜花”相关的内容推荐。

需要注意的是，这个阶段的数据是没有经过任何数据清洗的，因此可能会存在大量的非法数据。使用这些数据的下游应用应当具备处理非法数据的能力，或者能容忍非法数据导致的干扰。在上面的例子中，推荐模块主要关注的是一个人的近期搜索行为，而大部分非法数据并不是针对某个具体的个人，因此这里受到非法数据的影响较小。

数据清洗

在榜单计算的例子中我们介绍过，行为数据会存在大量各种行为的非法数据，因此数据清洗对于很多数据应用都是一个必要的步骤。不过，通常来说，除非是能 100% 确定的非法数据，例如没有通过基本的格式校验，否则数据清洗并不会真的完全过滤数据，而是通过对数据附加一系列的标签来标注数据是否非法，例如是否搜索引擎、是否是黑名单 IP 等。其中，部分标签甚至是一个概率值，例如数据清洗模块会标注某条行为数据有 80% 的概率是非人类访问。这些标签从不同的角度描述了一条行为数据的可靠性，下游的应用可以根据各种业务需要，利用这些标签的值来灵活决定是否需要信任某一条行为数据。

由于实时数据清洗的局限性，多数场景下我们还需要定期进行离线的数据清洗，以便进一步标记出非法数据，提高数据准确性。和实时清洗类似，离线数据清洗依然采用标签的方式进行，这些标签可以复用。

数据存储

所有的行为数据最终都需要落到一个持久化的、高效访问的存储系统中。一般来说，像 HDFS 这样的分布式文件系统是一个很好的选择。

但是，正如我们在客服系统中的例子里提到的，不同的数据应用可能需要不同的数据查询访问，HDFS 虽然是一个可靠的存储，但是显然对于根据 ID 进行随机查询这件事情非常不擅长。因此，通常一个数据平台的存储体系都是多个组件相结合，例如以 HDFS 来提供高吞吐的基础存储，以 HBase 来提供随机更新和查询的能力。

数据计算

数据计算是一个更复杂的话题，类似于数据存储，不同的数据应用对数据计算的需求也是各不相同的。前文提到的榜单计算中，就同时需要实时计算和批量计算两种不同的模式。因此，一个数据平台通常应该具备多种计算能力，例如提供以 MapReduce、Spark 为代表的批量计算框架，以及以 Storm、Spark Streaming 为代表的流式计算框架。

API 查询

虽然数据存储层已经提供了完全的数据访问能力，但是底层的接口往往是复杂而难以使用的，并且使用不当可能会导致资源的浪费。因此，对于常见的查询场景，数据平台通常还需要提供一套查询 API，把底层的查询能力进行封装。

上文我们提到客服系统中需要根据用户 ID 进行用户行为序列的查询，这是一个典型的 API 查询的使用场景。同样的 API 还可以被直接用在浏览轨迹、用户行为调研等其他数据应用中。

SQL 查询

SQL 提供了高效、灵活的数据处理能力，在数据调研、数据预处理等阶段都是非常便利的手段。因此，数据平台必须要提供针对用户行为数据的 SQL 访问接口。

由于用户行为数据的数据量通常是非常大的，我们不建议专门针对 SQL 访问的需求单独进行数据存储，以免导致不必要的计算和存储成本。更合适的方式是基于已有的数据存储，配合合适的 SQL 查询引擎，直接进行数据查询。

具体来说，常见的 SQL 需求可以分为两大类。第一类是统计分析类，例如 PV、UV 等汇总类查询。这类查询的特点是会产生一个较小的结果集，但是对响

应时间的要求比较高，例如秒级或者分钟级。第二类是数据处理类，例如格式转换、条件抽取等常见的数据预处理操作。这类查询的特点是产生结果的大小和原始数据集相当，甚至有可能会更大（例如JOIN）。

在超大规模数据量的平台体系下，针对上述两类需求有可能需要采用不同的SQL引擎来提供支持。因此，数据平台中通常也需要支持多个SQL查询引擎，以便在合适的应用上选择合适的查询引擎来使用。

总结

用户行为数据可以被应用在产品、服务的各个层面，而底层的数据平台则是这一切的基础。只有构建了一个灵活、易用的数据平台，才能实现高效的数据应用迭代，让数据更快地对产品和服务产生价值，最终实现真正意义上的数据驱动业务。

数据应用与用户智能

有了数据平台，我们就有了数据根基和数据处理能力，接下来就是围绕数据的产品智能应用。具体到用户行为数据，它不仅可用作流量统计和在线的用户行为分析，还可以应用于各种用户智能。图5-7是挖掘用户行为数据价值的一系列智能应用。

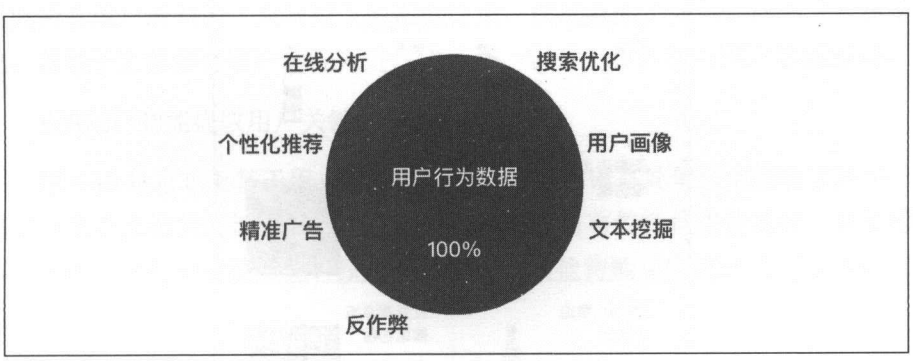


图5-7　挖掘用户行为数据价值的一系列智能应用

基于用户行为数据的用户智能应用

目前常见的基于用户行为的用户智能应用，如表5-1所示。

表 5-1 常见的用户智能应用

应用类型	解决的问题
用户画像	描绘目标用户到底是一个什么样的人——目标用户的属性、行为与期待等
个性化推荐	基于用户画像、短期兴趣、长期爱好，向用户推荐喜欢的内容，如视频等
精准广告推荐	根据用户以往的浏览和购买等行为，向用户展现最有可能感兴趣的广告
精准用户运营	根据用户以往的浏览和购买等行为，向用户推送最有可能转化的优惠券等
反作弊分析	判断用户是否是一个“作弊”用户
搜索引擎点击模型	基于海量用户检索的点击行为，调整检索结果中高频点击项的排序
智能评价系统	自动从用户评价中，抽取关键字及情绪化文字，如“大屏幕”、“超长待机”、“老人家喜欢”等
流失用户预警	提前预警用户潜在的流失倾向，提供优惠券、促销活动，延长用户生命周期
导航行程时间预估	根据天气情况、实时路况信息、海量用户历史行程记录，预测导航行程时间

下面列举我们常见的用户智能的应用。

应用 1：个性化内容推荐

图 5-8 是一个电商 APP 的首页，其中每一个商品，都是根据用户以往的浏览与购买记录做出的具体的个性化推荐，不同的用户，打开 App 后看到的都是不同的内容，真正做到了“千人千面”。

图 5-8　某电商网站首页的智能推荐（图片来源于网络）

应用 2：根据搜索结果的页面排序调整

图 5-9 是百度的搜索结果页，每个用户看到的搜索结果一样，然而这些搜索结果的具体排序，是基于不同用户对同一个搜索词的点击情况进行调整的。

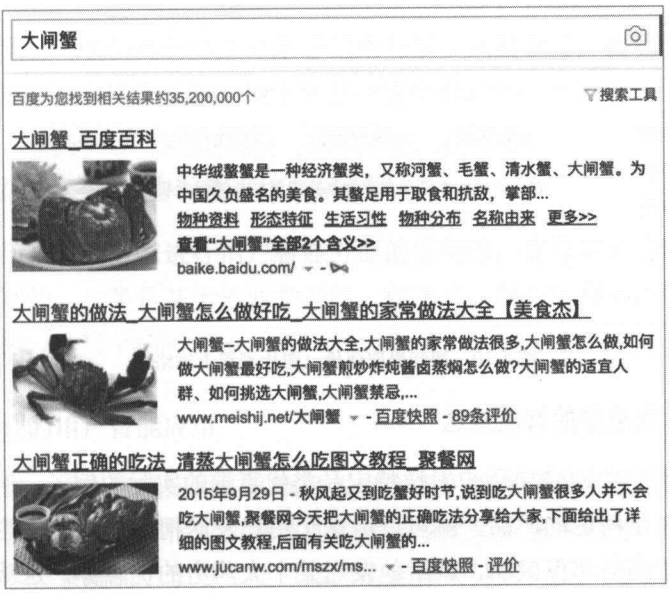


图 5-9　百度搜索“大闸蟹”结果页（图片来源于网络）

例如，在搜索“大闸蟹”的结果页中，如果用户点击搜索结果中第二条的比例要远远高于点击第一条的比例，那么第二条的顺序就调整到第一条之前，因此后面所有用户看到的“大闸蟹”的搜索结果，就都发生了变化。简言之，这个变化，是基于之前所有用户的行为进行的调整，也是一种典型的用户智能应用。

应用 3：智能提取用户关键性评价

图 5-10 是京东上基于用户的商品评价，用自然语言处理技术提取了其中一些具有典型意义的关键性评价。也只有对这些自然语言的评价进行这样一种处理之后，才能够进行后续的汇总、统计与分析，对其他待购买的用户具有参考意义。

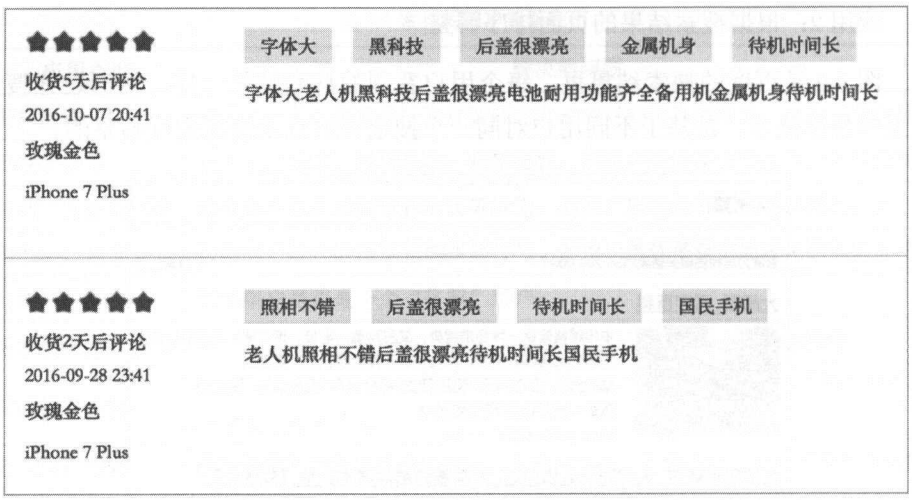


图 5-10　智能提取评价（图片来源于网络）

应用 4：优惠券的智能推送

图 5-11 是一个对特定用户发送特定品类优惠券的实际的例子，例如，一个用户以前经常浏览柯达的产品，就可以给这个用户发送柯达产品的优惠券；一个用户经常买一些药品和保健品，就给他发送某个大药房的优惠券。这就避免了给所有用户发送同样优惠券的粗放运营方式。在如今这个产品竞争激烈的年代，精细化运营是每一个企业战胜对手和最大限度挖掘用户价值的必备手段，也是用户智能应用的一种体现。

图 5-11　用户优惠券的智能推送（图片来源于网络）

当然，除了上述表格中记录的这些应用之外，用户行为数据在用户智能方面，还会有很多其他的应用，这里不再赘述。简单来说，只要是通过收集用户以往的数据，运用强大的工具和算法得出新的结论，创造新的知识，就可以被视为用户智能。

用户智能最常见的两种做法是基于规则与基于机器学习，被提到最多的两类应用是用户画像与个性化推荐。接下来会分别介绍这些内容。

用户智能分类：基于规则与机器学习

在基于用户行为数据进行用户智能方面的应用时，有很常见的两类方法，一类是基于规则的；一类是基于机器学习的。接下来，我们分别对这两类方法以及相关的应用做一个简单的阐述。

基于规则的用户智能应用

很多时候，用户智能应用的实现并没有想象中的那么难，在保证整个数据流完备的情况下，只要基于行业和领域的经验与专家知识基本规则，就能取得比较好的效果。这类系统的“学习”能力靠的是业务人员的头脑，为了便于理解，我们也姑且把这种场景也当作一种“智能”。我们以一个实际的案例来进行讲解。

这是一个电商网站，新上架了一款很好的智能手机，如图 5-12 所示，网站希望在特价时段将它推荐给最合适的用户。

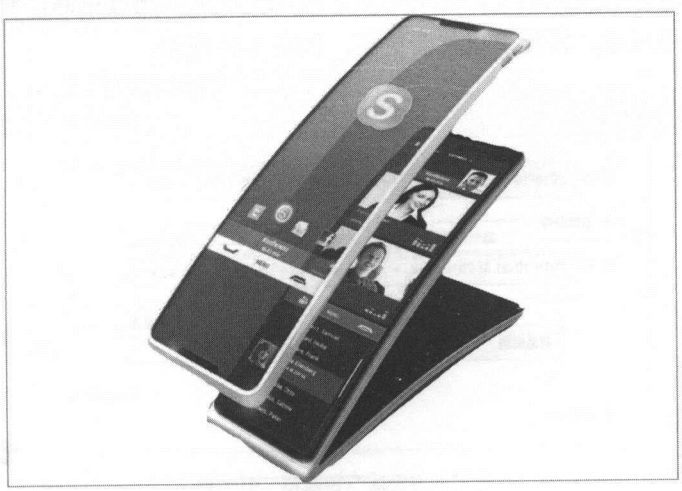


图 5-12　一款待推荐的智能手机（图片来源于网络）

为推荐这款手机，该网站的运营人员决定先向最近一个月内的活跃用户发送促销信息。他们使用类似于神策分析这样的用户行为分析工具，通过规则直接找到过去一个月活跃用户的用户 ID 明细，通过第三方或者自研的推送系统来给这些用户推送具体的新产品信息，如图 5-13 所示。

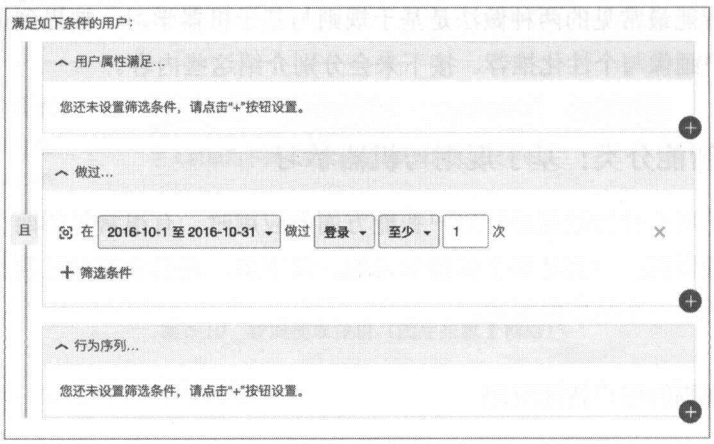


图 5-13　通过用户分群找到最近一个月的活跃用户

获取这些活跃用户之后，运营人员发现：用户数量太多，如果给所有人都发促销信息，就会给大量客户带来骚扰。毕竟这种推送带有浓重广告色彩，本身是极易产生负面影响的运营手段，不能滥用。

因此，运营人员进一步缩小了范围，只给在过去一个月中购买过电子产品的用户发送，因为购买过电子产品的用户，对电子产品兴趣度更高，对于推送的消息可能不会很反感，转化率也就会更高，如图 5-14 所示。

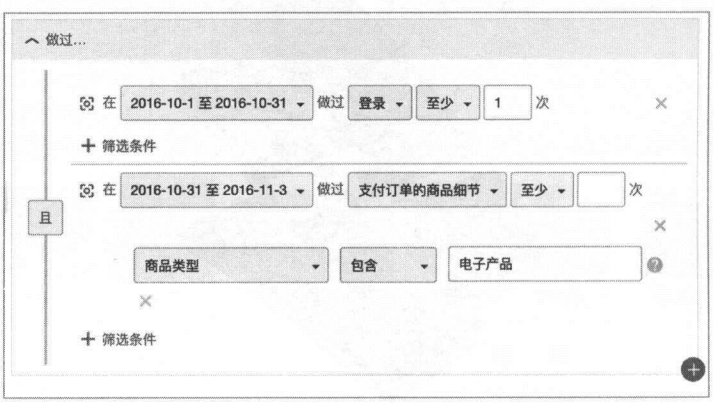


图 5-14　筛选过去一个月购买过电子产品的用户

如果运营人员觉得这还不够精准，可以筛选出注册比较久的老用户，对这部分群体发送促销信息，把它视作给这些资深用户的一个福利与回馈。可以进一步增加用户属性上的注册时间的限定规则，如图 5-15 所示。

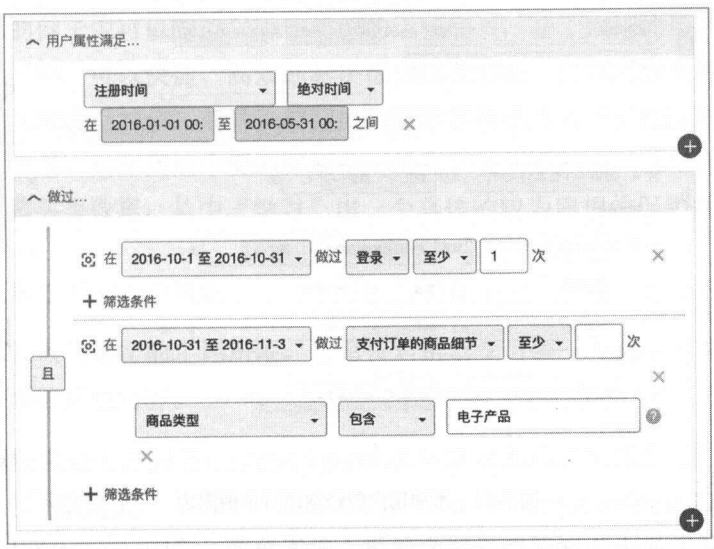


图 5-15　增加注册时间的限定

由于产品价格比较高，为保证最终的转化率，目标对象需要有一定的消费能力，则应该进一步筛选出这部分用户：限定上个月购买产品的价格，只对那些购买过高价值产品的人进行推送，如图 5-16 所示。

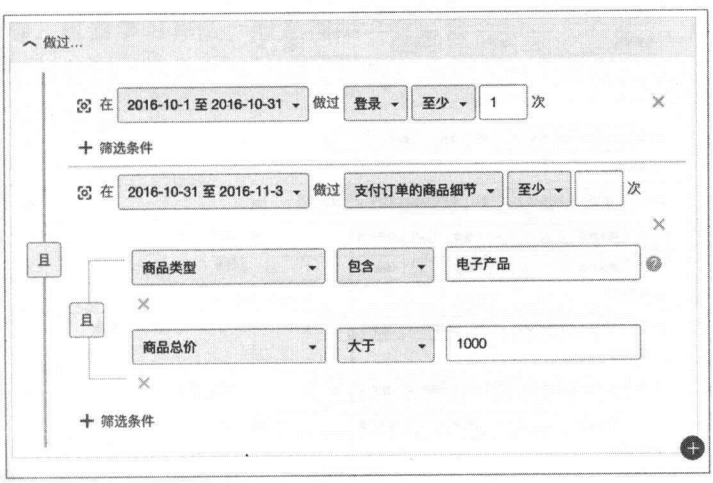


图 5-16　增加对产品价格的限定

同样，我们还可以继续增加更多的限定规则。例如，我们只对那些有过取消订单操作的客户发送这个促销消息，如图 5-17 所示。

图 5-17　增加用户曾经取消订单的限定

或者可以再进一步增加浏览电子产品到一定次数，表现出明显兴趣倾向这一限定规则，如图 5-18 所示。

图 5-18　增加用户浏览商品次数的限定

只要是企业所选择的第三方数据分析产品本身能够满足，筛选规则可根据运营需求的精细化不停地添加，直到筛选出符合需求的那些用户群体，随后对他们进行推送。

从上面的例子中可以看出，对于类似的运营场景，基于规则的用户智能已经非常复杂和精细，可以快捷、高效地解决很多运营需求。但是，如果规则没有办法很好地由人来进行抽象，或者说规则已经复杂到没法由人来描述，就需要用到机器学习的方案。具体到这个促销案例中，机器学习所要做的，就是根据以往用户的浏览、购买等数据，从中“学习”出一个具体的较为理想的促销规则。

基于机器学习的用户智能

由于篇幅以及本文的主题所限，这里仅对机器学习做一个简单的介绍，便于读者理解后面的其他内容。

人工智能领域的先驱者，Arthur Samuel 在 1959 年创造“机器学习”这个概念时，这样对它下的定义：“Field of study that gives computers the ability to learn without being explicitly programmed”。简单来说，机器学习是研究通过不显式编程来赋予计算机学习能力的一个领域。从这个概念可以看出，与以往普通的计算机程序有一个最大的不同，就是机器学习并不是一个被完全设计好的程序，而是一种特殊的、能够自我提升的算法，让计算机自己从数据中学习并由此具备解决问题的能力。

机器学习的常见算法有很多，如常见的回归算法、分类算法、聚类算法、关联分析算法等。机器学习算法一般是按照“有监督”和“无监督”，以及模型最终输出目标是映射到“连续”空间和“离散”空间这两个维度来分类，划分的 4 个类别如图 5-19 所示。

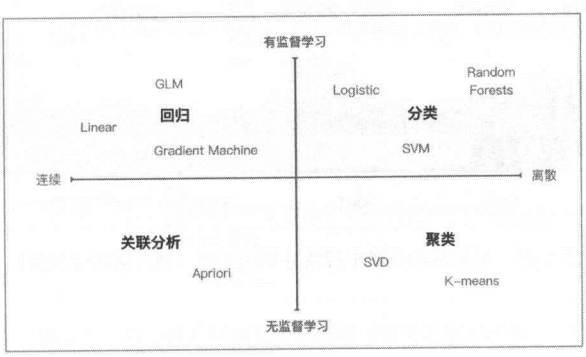


图 5-19　常见机器学习算法的分类

我们先来大概描述一下什么是“有监督”学习和“无监督”学习，简单来说，是否“有监督”，就看输入给机器的那些数据，是有标签的还是没有标签的。“有监督”学习就意味着提交的不仅仅有“问题”，还有“答案”。而算法是“连续”还是“离散”，则主要取决于算法输出的结果是无限多个还是有限多个。我们从这个角度来看看几种常见的机器学习算法。

1．回归算法

回归算法是一个典型的“有监督”学习算法，因为它的输入是典型的有目标值的。同时，回归算法拟合的结果，一般而言都是连续值，所以它是典型的“有监督”的“连续”算法。

回归算法在很多领域得到了普遍应用，在金融方面，可以用它来做股市行情分析和预测，如图 5-20 所示；在产品运营方面，可以用它来做产品流量预估；在生物领域，可以用它来做蛋白结合点位预测；在交通领域，可以用它来做道路流量预警。

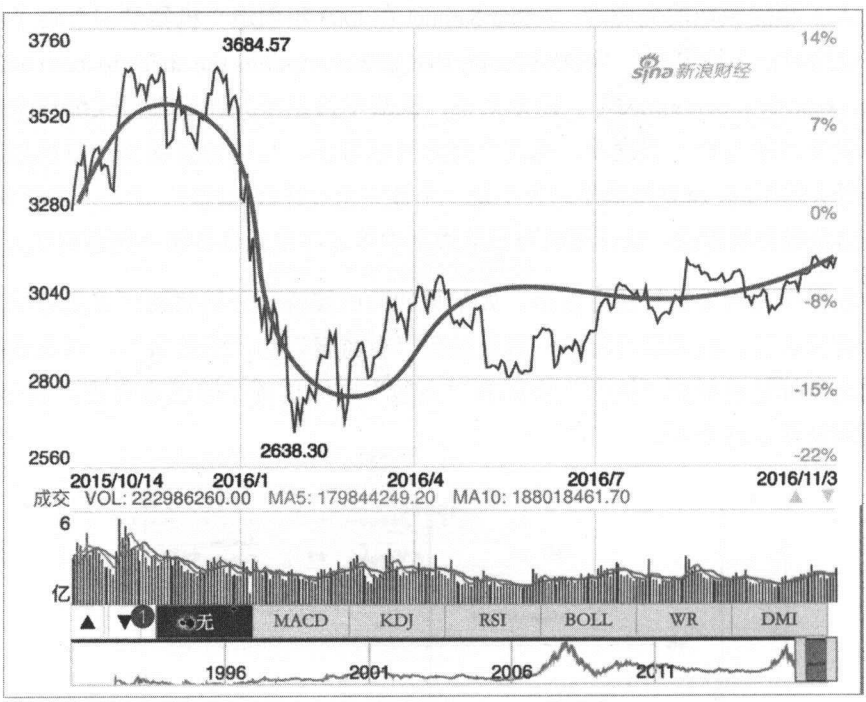


图 5-20　回归算法在股市行情分析和预测（图片来源于网络）

2．分类算法

分类算法是另一种经典的“有监督”算法，它的训练集里面，每一个样本都

是有确切的分类 tag 作为“正确答案”的。同时，它也是一个“离散”算法，因为它的输出结果是一系列的分类 tag，是一个个的离散值。SVM（Support Vector Machine，支持向量机）是一种分类算法，图 5-21 是用 SVM 对二维平面上的样本进行分类。分类算法也在各个领域都得到了广泛的应用，在金融方面，可以用它来识别作弊用户；在交通领域，车牌识别也是一种分类应用；在产品运营领域，它可以用作流失客户的预警，提前找到那些有可能流失的客户。

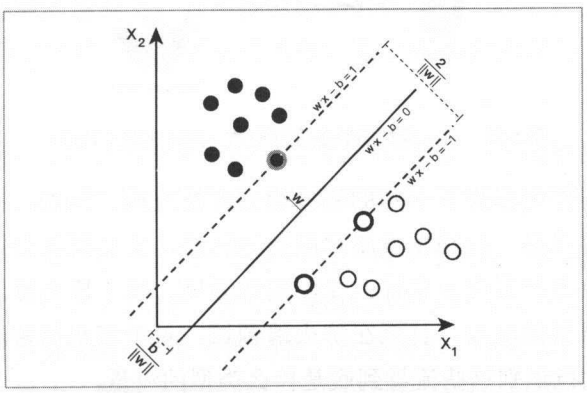


图 5-21　用 SVM 对二维平面上的样本进行分类

3．聚类算法

与分类算法不同，聚类算法是一个典型的“无监督”算法。聚类没有训练样本这一概念，所要做的就是基于输入样本的某些特征，按照一定的评价体系，将它们按照相似程度聚成几类。以图 5-22 和图 5-23 为例，同样的一些头像，按照聚类的标准，可以聚类得到不同的结果。

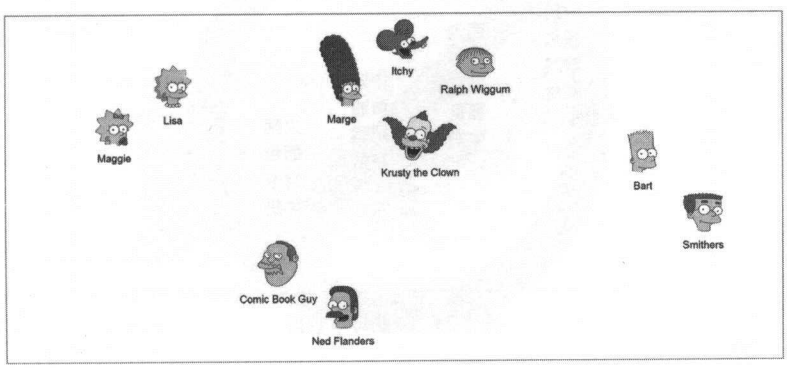


图 5-22　对头像按照发型进行聚类（图片来源于网络）

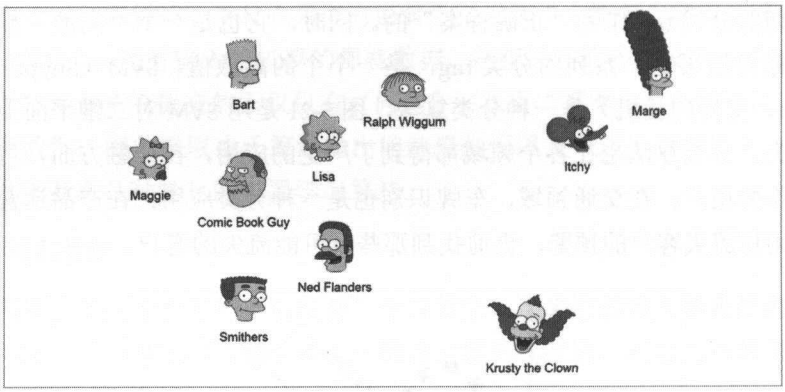


图 5-23　对头像按照肤色进行聚类（图片来源于网络）

聚类算法同样也在很多不同的领域得到广泛应用。例如，在产品运营中，我们可以用聚类算法，自动聚合用户的行为轨迹，并且据此分析用户使用产品的习惯，用于改善产品交互设计；在图形学领域，基于聚类算法进行图像主题筛选也是一种常见的应用；甚至在防火墙领域，对于那些加密流量，也可以通过聚类算法来大概识别这些流量到底是什么类型的流量。

4．关联分析

最后一种典型的“无监督”的“连续”的机器学习算法，我们就以关联分析为例来进行介绍。关联分析的一个非常典型的应用，是从订单数据中分析两个商品的关系。如图 5-24 是一些订单数据。

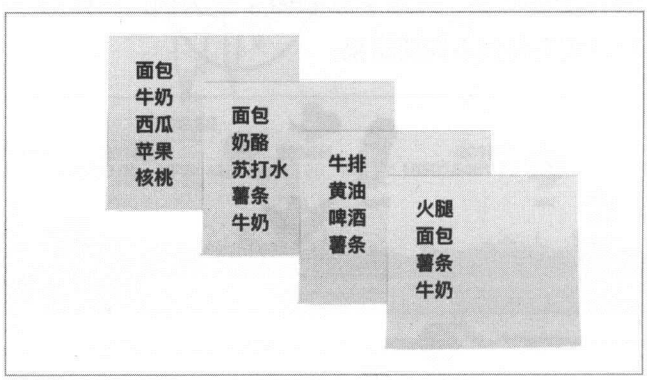


图 5-24　一些订单数据

我们想从这些订单数据中分析商品之间的关系，可以基于关联分析（有时候也被叫作关联规则挖掘）来做。

我们可以分别计算两个商品的支持度（Support）与置信度（Confidence）。例如在超市购买(D)这个事件中，包含了买了啤酒(A)的事件，买了面包(B)的事件，以及同时买了啤酒和面包(AB)的事件。支持度的定义是：Support(A -> B)=P(AB)，它揭示了A与B同时出现的概率。如果A与B同时出现的概率小，说明A与B的关系不大；如果A与B频繁同时出现，则说明A与B总是相关的。置信度的公式是：Confidence(A -> B)=P(B|A)，它揭示了事件A成立的条件下，事件B出现的概率，比如在买啤酒的情况下，有多少概率会去买面包。如果置信度度为100%，则A和B可以捆绑销售，如果置信度太低，则说明A的出现与B是否出现关系不大。

基于上面的规则得到的支持度和置信度，可以搭建一个简单的商品推荐系统，在实际应用中，如相关电影推荐等应用，就可以用关联分析得到很好的解决。

接下来介绍关联分析和协同过滤的区别。协同过滤是一种间接推荐，即先找到品味相似的人（User Based），然后再根据品味相似的人的偏好进行推荐。适用于重度个性化并且Item[1]非常多的场景，比如音乐，电影等。关联规则是更加直接的推荐，从整体的数据中挖掘商品之间的潜在关联，与单个人的偏好无关，适用于Item不多，并且非重度个性化的场景，如超市购物、汽车导购、交通规划等。

完成基本讲解后，我们再回到前面的促销案例。现在我们不再局限于基于规则的促销方法，而是尝试用机器学习来解决这个问题，如图5-25所示。

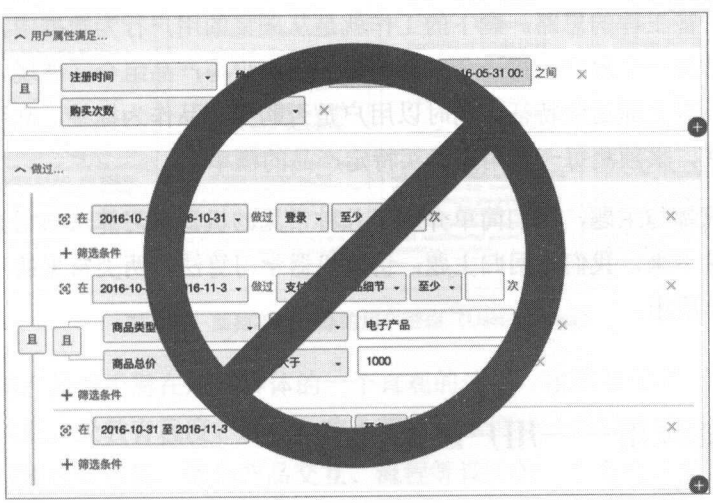


图5-25　不再基于规则而基于机器学习

1　Item，（被推荐）物品。

通常，一个典型的机器学习处理流程会包括问题分析、数据清洗、特征工程、模型训练、模型验证。

我们用回归算法来解决这个问题，将问题重新定义为“预测用户购买促销产品的概率”，并据此给概率最高的用户发送促销信息。

对于机器学习算法而言，数据清洗与特征工程是最重要的一部分工作。通过特征工程，我们可以将对解决问题有意义的关键信息从海量数据中抽取出来，让数据中的“规则”浮现，同时结合自己的先验知识更有效地进行处理和分析。

对于预测产品购买概率的问题，根据业务模式的先验知识，我们可以从用户的兴趣、产品的受欢迎程度和用户对该产品的兴趣来提取特征。

实际上，把一些特征组合起来可能会比单个特征提供更多的信息，甚至还可以对他们进行加减乘除等操作来获得更有效的特征。例如产品在最近 7 天内的购买率为购买次数 / 浏览次数。

此外，还可以利用不同的方法表示数据，以获得更多的信息。例如将购买次数映射到极少（0 ～ 10）、少（10 ～ 100）、中等（100 ～ 1000）、多（1000+）等分类中。

特征工程有无穷的可能，以上只是针对产品购买率预测这一个问题，提出了最常见的特征工程的思路。剩下的工作就是从采集的用户行为数据中，按照上面的思路，对每一个用户、每一个产品、每一个用户与产品组合（展现、浏览、购买），都抽取上面这些特征，同时以用户是否购买产品作为标签，套用经典的逻辑回归算法，来预测每一个用户购买特定产品的概率。

限于篇幅与主题，我们简单介绍了几种常见的机器学习算法与它们的一些典型应用。接下来，我们将回归主题，讲解机器学习算法在两类常见的用户智能应用中的具体应用。

用户智能应用——用户画像

两种用户画像：User Persona 与 User Profile

很多企业，都有建设“用户画像”的需求。首先来介绍我们所理解的两种用

户画像（User Persona 和 User Profile），以及如何构建用户画像（User Profile）的标签体系，并由此驱动产品智能。

User Persona

第一种意义上的用户画像（User Persona），是产品设计、运营人员从用户群体中抽象出来的典型用户。例如，在用户调研阶段，产品经理经过调查问卷、客户访谈了解用户的共性与差异，汇总成不同的虚拟用户；在产品原型设计、开发阶段，产品经理围绕这些虚拟用户的需求、场景，研究设计产品用户体验与使用流程；当产品设计出现分歧时，产品经理能够借助用户画像（User Persona）跳出离散的需求，聚焦到目标用户，不再讨论这个功能到底要不要保留，而是讨论用户可能需要这个功能，如何使用这个功能等等。

图 5-26 是某招聘类产品（https://www.clearvoice.com）在调研阶段构建的用户画像（User Persona）。

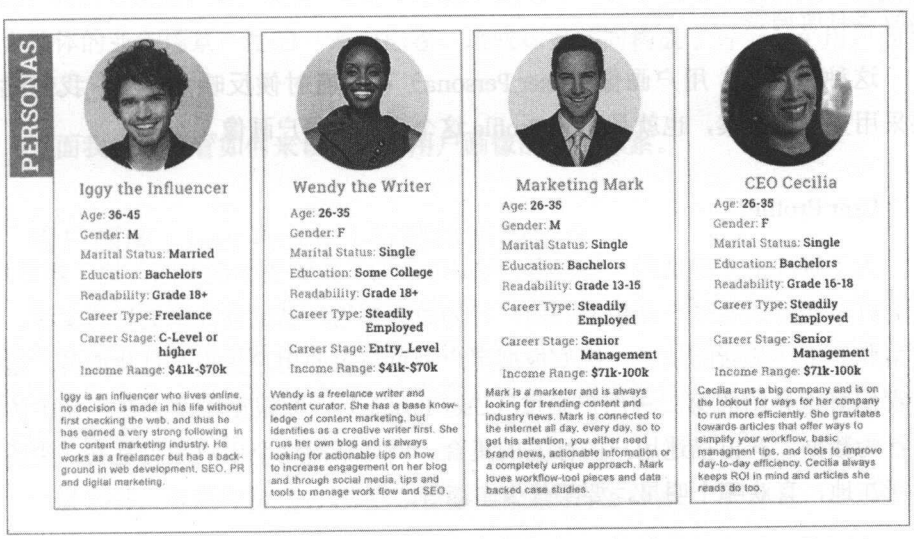


图 5-26 某招聘类产品的用户画像（User Persona）

这是该产品对于潜在用户群体的一个直观的认识，包括这些用户的年龄、性别、婚姻状况、受教育程度、职业、收入等各个维度的一些预估。而后续这个产品就可以按照这些预估，作为产品交互、流程等设计的一个重要依据。

所以，总的来说，这类用户画像（User Persona）本质上是一个用来描述用户需求的方法论。它可以帮助不同角色在产品研发过程中，从用户的角度思考

问题。在产品设计阶段和原型开发阶段，产品经理会较多地借助用户画像（User Persona）理解用户的需求，想象用户使用的场景。

但在通常情况下，随着产品上线后不断迭代，真实用户越来越多，仅通过用户画像（User Persona）可能难以更加量化细致地评估用户需求，也很难通过数据来确定用户画像（User Persona）虚构的人物是不是真的目标群体。同时，真实用户群体也随时间推移变化，在设计阶段虚构的用户画像（User Persona）需要重新调研、设想。

新浪微博就是一个典型的例子，最开始，微博的设计主要是为了满足一二线城市白领的使用，此时的用户画像（User Persona）可能是这样的：一二线城市、二十岁到三十岁、较高教育程度、白领、收入在 6000 元 / 月以上，这时，新浪微博所有的产品交互和流程设计可能都是据此进行的。但是，随着新浪微博的逐渐发展，它的用户群体已经发生了明显的“下沉”，越来越多三四线城市的“草根”用户开始使用。那么，对于这种情况，新浪微博的整个产品、功能、交互设计等，都应该有所调整。

这种情况下，用户画像（User Persona）可能有时候反映会滞后，我们就需要采用另一种方案，也就是 User Profile 这个意义的用户画像。

User Profile

为了解决上文提到的一些问题，同时也是为了能够更加精细深入地了解用户，我们自然会希望通过产品积累的用户行为数据来为产品运营提供更好的支撑，甚至由此诞生一些新的功能，例如根据用户浏览记录向用户提供个性化服务。这就是我们着重介绍的第二种用户画像（User Profile），即根据每个人在产品中的用户行为数据，产出描述用户的标签的集合。例如猜测这个用户是男是女，生活工作所在地，喜欢哪个明星，要买什么东西等。

特别是随着“千人千面”等理念深入人心，许多企业希望能建立自己的用户画像体系。那么，在这种情况下，我们更应该明确两种用户画像的差异。与第一种用户画像（User Persona）不同的是，用户画像（User Profile）的建设更加关注以下几个方面。

- 是否反映受众的真实需求：用户画像（User Profile）这个词的字面意义，是关注人口属性、生活状态、人生阶段等静态信息，但这些信息并不一定直接反

映用户兴趣。产品更关注的往往是某用户“最近喜欢看哪类视频”、“准备买多少钱的手机”这些能够帮助产品运营的动态信息。

• 时效性：用户的兴趣偏好随时都在发生变化，需要及时更新用户标签。极端情况下，我们甚至希望用户上一次浏览的情况，在他进行下一次浏览前就能体现并更新到用户画像（User Profile）上。

• 覆盖度：用户画像（User Profile）既要勾勒出用户感兴趣的内容，也要记录用户不感兴趣的信息，尽量多地满足产品运营的需要。但同时，除了人口属性等明确的属性外，大多数用户画像的正确与否是没有意义的。如“最近喜欢看搞笑视频”这个标签，并不表示用户下一次一定观看搞笑视频，因此执着于提升标签的准确度，不如设计出更多清晰描述受众需求的标签，更多时候我们注重提升用户画像的覆盖度，同时提供更细粒度的画像。

简而言之，用户画像（User Persona）主要来源于产品与运营人员对客户的理解、调研与认知，用户画像（User Profile）则主要是基于真实积累的用户行为，结合具体的业务场景产生的一系列标签，这些标签共同构成了对于一个用户的真实描述。

下面我们就看看如何来设计一套用户画像的标签体系。

用户画像（User Profile）标签体系的建立

所谓用户画像（以下均指 User Profile）中的标签体系，简单来说就是将用户划分到多少个不同的分类之中。当然，在这种情况下，一个用户是可以归到多个不同的分类上的。用户落入的这些分类都是什么，彼此之间有何联系，就构成了一个标签体系。

一般来讲，有两种常见的思路设计用户画像的标签体系。

一类是结构化的标签体系，这类标签可以直接从人口属性、物品信息等基本信息中直接得到，有明确的层级关系，如性别、省市、视频分类、商品分类等。图 5-27 是亚马逊（http://www.amazon.cn）的商品标签体系，用户画像的标签体系与此类似，可以结合具体的业务场景来确定。

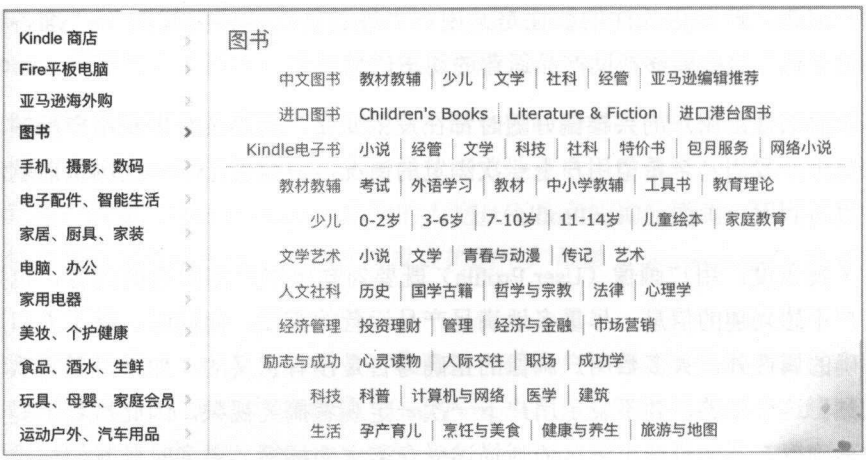


图 5-27　亚马逊的商品标签体系

简单来说，结构化的标签体系通常较为简单，一般可以直接通过用户的行为映射得到。例如根据用户的购买记录，为用户构建物品对应的结构化标签。但结构化标签往往粒度较粗，无法充分衡量用户的兴趣，例如用户在新闻类 APP 中阅读了关于某明星的娱乐类新闻，并无法推断出他对所有娱乐类新闻感兴趣，他也不一定只对该明星情有独钟。

另一种是非结构化标签体系，就是各个标签各自反映各自的用户兴趣，彼此之间并无层级关系。典型的非结构化的标签，如搜索广告系统中的关键词，或者文档主题模型（Topic Model）。例如新闻类 APP 中，我们往往会构建大规模的主题模型（主题数在千万级别），不仅仅涵盖已经构建的结构化的标签体系，如娱乐（明星、搞笑）、体育（篮球、足球）等，还能更细致地表达如星座、食物、体育活动等语义上的分类，而且这些分类之间并没有明显的层级关系。

标签体系的建设本身一要便于使用，二要有明显的区分度。结合具体的产品而言，在不同的场景下对这两点要求的核心是不同的。因为选择哪些标签并没有明确的依据，还是需要充分了解到底是什么因素在驱动用户使用产品。有效的标签体系，要能反映用户决定买什么、不买什么的逻辑与依据。例如电商产品中，以新闻频道的方法，为用户构建“财经”“体育”“旅游”等标签，虽然并不难，但也没多大意义。

实践案例

我们曾经与国内某知名视频聚合网站共同搭建视频推荐服务。该网站每天聚

合全网的视频，向用户提供热门视频、视频检索等服务。网站已经积累了大量的用户和行为数据，围绕新、老用户的运营模型发生着变化。

在开始具体的项目之前，我们首先要意识到，与传统的视频站点不同，短视频网站是有它自己的运营特点的，这些特点包括以下几点。

1．播放随意性强。短视频播放虽然是个高频、周期性强的用户行为，但单次观影时间短，用户选择随意性大。

2．热点轮换迅速。平台中不断加入新视频，每天的热门内容不断变化，网站需要发现用户潜在的兴趣点，向用户推荐新鲜内容。

3．场景驱动。场景是特定的时间、地点和人物的组合下的特定的消费意图。不同的时间、地点，不同类型的用户的消费意图会有差异。例如白领乘地铁上班，会关注当日的新闻热点；周末晚上在家，用户更喜欢点击娱乐类搞笑视频。当场景辨识越细致，就越能了解用户的消费意图，推荐的满意度也就越高。

随着视频资源的不断丰富和用户需求的多样化，如何准确地向客户推荐视频，是该产品用户画像的一个基本目标。我们十分看重推荐系统中推荐结果的可解释性，也让用户能感觉到每一条推荐视频的推荐理由。当然，我们构建用户画像也以观看场景和观看兴趣为主。

我们考虑新用户和老用户两大类群体。新用户第一次进入 APP，在这一阶段的运营目标以留存为主，主要向用户推荐近期热门视频。除了常规的设备信息、地理信息外，我们对用户了解甚少，可以通过猜测“用户在哪里”、“这个时段可能处于什么场景”来构建用户画像，进行场景推荐。这两种标签的获取较为直接，通过用户手机的地理位置信息和当前时段就可以得到。

基于这两个标签，在不同场景下，我们向新用户推荐不同的视频，例如：

工作日 7:00 ～ 10:00：用户可能搭乘公共交通工具前往公司，乘车时使用 3G/4G 流量上网，时间较为碎片化，并且容易受到打扰而中断观看。通常这个时段用户希望了解当天的时事、新闻。因此我们推荐短小精悍的热点新闻。

工作日 12:30 ～ 14:00：用户可能在公司午休，我们推荐娱乐、搞笑类的视频，目的性较弱，随意寻找符合自己口味的内容，但有可能因为午睡或工作，观影时间碎片化。因此，我们推荐视频时长较短，诸如娱乐、搞笑类的视频。

周末 19:00 ～ 23:00：用户可能在家中休息，观看时间较为充足，并且使用

WiFi，速度稳定。这个时段用户目的性通常较强，例如观看“XX 歌手”、“XX 男”等综艺节目的热门片段。因此我们可以推荐综艺节目、电影片花等，满足用户长时间放松的需求。

通过场景推荐的方式，我们可以在不了解用户兴趣的情况下，针对不同场景标签下的新用户推荐不同热门视频，满足用户需求。

而对于老用户，运营目标是提升用户体验，向用户推荐感兴趣的内容，以提高观影时长；结合场景推荐用户可能感兴趣的新鲜内容，以提高用户留存率。除常规信息、场景信息外，构建老用户的用户画像还会考虑用户在不同时段的兴趣点、用户是否喜欢探索新鲜视频，以及对用户召回需求。下面我们分别对这三类进行描述。

对于第一类“用户兴趣标签”，我们可以通过视频本身的分类信息构建结构化的兴趣标签。在实际处理中，我们将每个用户最近观看记录作为一个观影序列，通过 Item 2 Vec[1] 产出视频的 Embedding[2] 矩阵，并用 Bag of Words 的思想以每个用户的最近观看记录描述用户兴趣，得到用户 Embedding，作为用户兴趣标签。通过用户兴趣标签，我们可以将用户兴趣融入前文描述的场景推荐中，例如在工作日的 7:00 ～ 10:00，我们根据用户兴趣，从热点新闻中筛选用户感兴趣的军事、财经等品类；在周末的 19:00 ～ 23:00，我们根据用户上周的观影记录，推荐新一期的综艺类节目。

对于第二类“用户新鲜度的需求标签”，我们通过衡量用户观影记录中各影片之间的相似度得到。影片分类覆盖越多，或影片之间的向量距离越远，说明用户越喜欢探索新内容。对于喜欢探索不同类型的视频的用户，我们会更倾向于从用户未观看过的分类中，抽取新鲜热门视频加入推荐排序结果。

对于第三类用户召回方面的需求，其实也是一个非常现实的需求。神策数据可以通过多维分析的方式寻找用户流失的原因，同样，我们也通过统计方法预测用户流失风险。例如，对于视频网站的老用户，观影习惯和场景通常较为固定，当用户最近一段时间内的观看频次显著低于过往，甚至没有打开 APP 时，我们判定用户有流失风险，可以通过推送感兴趣的视频等手段，召回用户。

---

[1] 《Item2Vec: Neural Item Embedding for Collaborative Filtering》，https://arxiv.org/pdf/1603.04259v2.pdf。

[2] Embedding，是对特征进行固定长度的编码。例如，对词进行固定长度的编码，即“Word Embedding”。

现在，让我们总结一下，短视频是一个高频、随意性强的产品，用户的观看行为受时间、场地等场景因素影响较大，需要对用户在不同场景下的观看行为做深入了解，归纳不同场景下用户个体需求、群体需求的差异，针对不同场景制定相应的推荐策略，这也是我们选择场景作为短视频产品用户画像的突破口的原因。

同时我们在构建视频推荐的用户画像时还面临如下挑战。

1．数据稀疏性。个人的观看记录相对整体的覆盖度是十分低的，不同的个体间重合度也很低。我们需要从这些稀疏的数据中得到个体、群体的兴趣标签。

2．用户兴趣变化快。用户的兴趣点随时间、热点变化，用户观看了几次关于某明星的短视频，并不代表第二天或未来用户会对他感兴趣。我们需要分别构建用户短期、长期的兴趣标签。

3．场景识别难。目前我们的场景识别以时间段为主，未加入地理位置信息，而后者能显著提高细粒度场景识别的准确度。

最后，总结一下文中提到的两种用户画像。User Persona 可以帮助我们形象地了解目标用户的行为特征，作为我们判断用户需求的依据；User Profile 从用户行为中构建各种标签，在用户生命周期中不断刻画用户意图，辅助产品运营。

画像标签体系的建设是不断迭代的过程，例如视频产品中，新的视频、新的热门话题不断产生，不断地研究和调整也就必不可少。只有根据产品运营的目标，灵活调整标签体系，才能取得最好的效果。

用户智能应用——个性化推荐

我们先简单描述个性化推荐的基本概念，再结合一个具体的案例场景来描述典型的个性化推荐系统是如何构建的。

个性化推荐的概念

个性化推荐目前可以查到的最早的记录，是 1995 年 3 月，卡耐基·梅隆大学的罗伯特·阿姆斯特朗（Robert Armstrong）等人在美国人工智能协会上提出的个性化导航系统 Web Watcher；斯坦福大学的马可·巴拉巴诺维奇（Marko Balabanovic）等人在同一会议上推出的个性化推荐系统 —— LIRA。当然，在这

之前，传统的“非个性化”推荐，例如编辑精选、热门榜单、店长推荐、超市折扣货柜等，在不同的行业和领域，早就得到了广泛的应用。这些“非个性化”推荐，主要由领域专家结合领域知识和对整体用户群体的理解，来给全体用户的一个推荐。当然，这种推荐容易受到马太效应[1]影响，导致给所有人的推荐结果千篇一律，同时由于推荐的结果数量也有限，所以容易埋没优质的长尾素材。

随着科学技术的进步，特别是数据采集手段的丰富，在信息（例如用户行为数据）足够丰富（甚至过载）的情况下，可以考虑使用个性化推荐系统，来帮助用户快速发现对自己有价值的信息，做到推荐结果千人千面，提高获取信息的效率，解决“非个性化”推荐中的一些问题，从而最终达到提升用户体验的目的。

架构实现

我们以一个具体的短视频个性化推荐的实际案例，来看一下典型的个性化推荐系统是如何实现的，如图 5-28 所示。在这个案例中，客户使用了神策分析的数据采集方案来采集用户行为与用户 Profile 数据，并且推荐系统基于神策分析的相应 API 和 PaaS 类型接口，共用计算与存储资源。

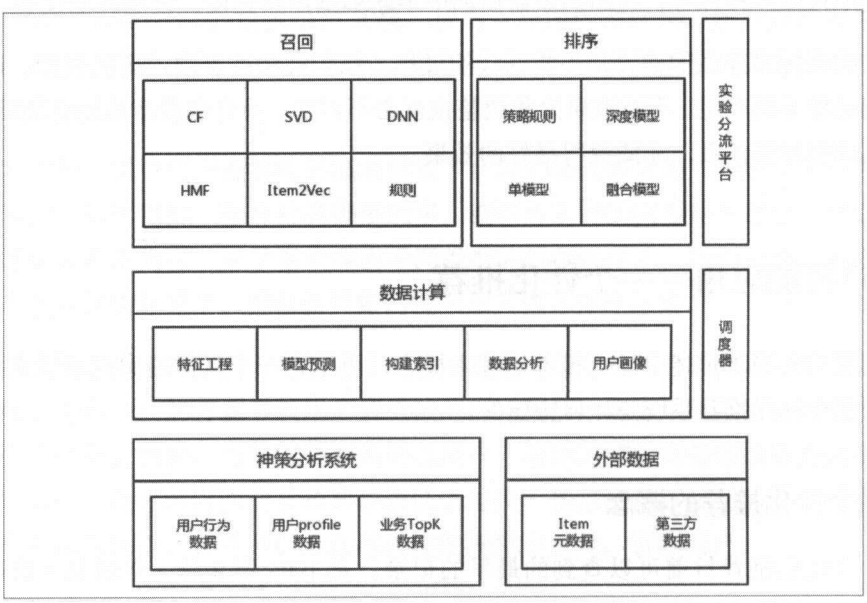


图 5-28　一个典型的个性化推荐系统的架构

[1] 马太效应（Matthew Effect），指强者愈强、弱者愈弱的现象，广泛应用于社会心理学、教育、金融以及科学领域。

下面，我们依次对架构图中的各个层级进行相应的说明。

首先是数据层。在这个项目中，个性化推荐所需的基础数据共由两部分组成，按来源可分为神策分析系统数据与外部数据。其中，用户的行为数据与用户的部分 Profile 数据天然地存储在神策分析系统中，这些数据格式规范，各业务间属性关联完备，为后续的数据计算工作打下了坚实的基础。但是，神策分析系统作为一款用户行为分析产品，目前本身并不持有客户的 Item 数据，所以需要额外提供推荐候选 Item 数据并商定其更新方式。而之前在个性化推荐业务上的一些积累，比如用户标签体系、分类体系等，也会对后续的数据建模等工作具有指导意义。另外，除了用户相关数据与 Item 相关数据，对于某些其他的应用，有时候也需要开发爬虫抓取第三方数据，例如，在进行小说推荐的时候，就可以通过爬虫抓取一些第三方网站的小说评价数据，当然，在这个短视频推荐的案例中，暂时没有用到爬虫。

上述的两部分数据有实时流计算和批量计算两种计算方式。其中，在实时流计算时，可直接从 kafka 订阅数据，而批量计算则是对已落盘至 HDFS 上的数据进行处理。

数据层之后是策略层。策略层本身又分为两层，一为基础数据计算层，紧随其后的是更加直接决定个推业务效果的召回与排序。由数据层汇总的数据，经过简单的 ETL，被抽象为 User、Item 和 Event 三张表，这些表构建起了整个短视频产品的数据集市。在此基础上，我们可以轻易地进行各项数据分析，特征工程，以及部分召回源的索引构建工作。

策略层中有召回和排序两个关键要素。

其中召回决定了对每一个用户的个性化推荐的候选短视频集合，它可能有多种来源与方法。例如，在这个案例中，我们已实现了诸如 CF（Collaborative Filtering，协同过滤）、SVD（Sigular Value Decomposition，奇异值分解）、HMF（hybrid matrix factorization，混合矩阵分解）、Item 2 Vec，当然还有最简单的人工规则与现在热议的 DNN（Deep Neural Networks，深度神经网络）等模型（方法）。同时，对于这些数据，我们根据产品的实际需求，分别按照优先级和具体比例完成了多召回源之间的组合，以便在确保召回源丰富的同时，控制合理的召回数量，保证前端在线服务的效率。

而个性化推荐的排序则比较复杂，其目标并不只是考察推荐结果的相关性，

还要看真实的业务场景与产品形态结合，以确定排序的目标。通常，我们采用CTR预估来融合各种召回策略得到的候选集，但在不同的业务场景中，需要不同的降权或过滤处理。例如，电商类的个推需要过滤已购买的Item，加强相关而非相似Item的推荐；而新闻视频类，则需要做内容去重，根据不同主题的特点，借助主题模型优化多样性或进行时效性加权。除了简单的策略规则排序，在这个案例的迭代过程中也尝试过单（机器学习）模型、融合模型（GBDT + LR，GBDT + FM）和深度模型。它们各有特色，单模型高效可解释性强，但特征工程较为复杂；深度模型端到端的特点，大大简化了特征工程的工作，但可解释性较差，不容易在模型上后续调优。在实际的应用中，需要针对具体需求来挑选最适合的模型。

对线上服务召回策略的变更与排序模型的修改，都需要通过实验分流平台来建立一组A/B测试。在这个案例中，这些A/B测试充分借助了神策分析系统的强大的分析能力，来实时对比流量效果，灵活响应，靠数据驱动来迭代业务升级。而最终的推荐Item，也都记录了其在产品中的路径，可以追溯其召回来源。

数据流

上一部分我们介绍了一个推荐系统的具体架构实现，这一部分我们将对系统的数据流进行具体介绍。图5-29是个性化推荐系统的完整数据流。

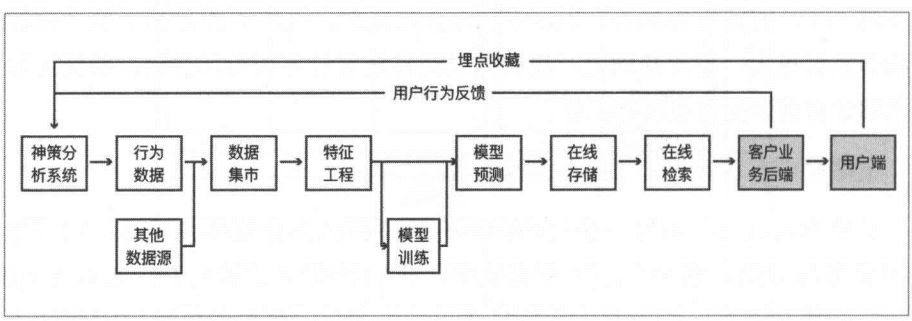


图5-29　个性化推荐系统的数据流

在这个数据流里面，神策推荐与神策分析系统在数据上形成了一个完美的闭环。神策分析系统可以提取用户行为数据，并结合其他数据源来构建基本的数据仓库。接下来通过特征工程（数据预处理、特征处理），完成数据的清洗、过滤、字典化、归一化、Embedding等，来产出适合模型读入的数据。经过模型训练与预测，便产生了个推的候选。其后，个性化推荐候选数据会被发送到推荐服务前

端机器上生成索引，经过压缩，灌入在线存储，从而通过推荐服务完成推荐业务。推荐结果在客户的用户端展现，新的用户行为则通过神策分析的数据采集系统，再次进入神策分析系统，作为新的数据参与后续的策略评估与修正，从而形成一个完整的闭环。

简而言之，神策分析系统天然地保存了用户行为数据及部分用户属性数据，只需要客户额外提供最基本的 Item 元数据，即可开始搭建最基本的个推业务。同时，使用神策分析平台，可以实时进行个推系统的效果查看，灵活分析不同实验分流在多维度指标上的表现，快速指导迭代决策。数据埋点作为推荐系统的基础，怎样才能支持好实验的灵活响应，亦是门学问。神策推荐可以免去客户自己在推荐服务前端的埋点工作。神策分析与神策推荐两者的深度融合，形成了数据流上完美的闭环，大力推进产品智能的进程。

业务分析与模型选择

在分别介绍完系统架构和数据流之后，本节我们将针对具体的业务分析与模型进行相应的介绍。

我们在着手准备推荐业务之前，首先需要对其现有业务有一定的理解。以短视频推荐案例为例，我们通过神策分析对这个短视频产品的业务数据做了一些简单的分析，用于指导我们后续的策略研发。这些分析指标包括活跃用户量、视频量、视频平均播放次数、视频平均观看时长等。从这个分析中，我们得到了一个初步的结论，对比每日众多的活跃用户量，相当比例的视频的播放次数非常有限，为长尾冷门视频。因此，在进行个性化推荐时，我们会尝试激活其中的高质量视频，同时也会挖掘热门视频，吸引用户观看产生更多的行为，以便后续业务迭代升级。与此同时，在这个分析过程中，我们也确定了此次个性化推荐的评价体系，也是日后迭代优化的目标，即从视频平均观看时长、用户留存、视频播放 CTR（Click Through Rate，点击通过率）这几个指标来衡量。

其次，我们进一步来分析数据的特点，以便协助进行模型的选择。对这个短视频推荐案例来说，它的用户行为数据量级足够大，每天会产生巨量的播放、点击行为；可推荐视频总量相比用户行为数据要小一个量级，并且已有一套自己的视频分类体系。基于上面的数据特点，我们决定以用户行为推荐为主，基于内容的推荐为辅。选择在深宽模型模式上采用 HMF 模型来生成候选集合，再通过主

题模型对推荐结果进行多样性优化（打散），最终辅以部分人工策略召回来构成我们的召回候选集合。

下面，我们对这三类模型做一个简单的介绍。

1．HMF 混合矩阵分解，即使用隐式反馈来做矩阵分解。隐式反馈多为用户正常使用产品所产生的行为，并非为了表达兴趣、态度，例如点击、播放、浏览详情页等。显式反馈则相反，例如评分、赞同 / 反对。我们采用隐式反馈，一来数据比显式反馈更加稠密，二来隐式反馈更代表用户的真实想法，三来它更容易激活一些小众的 Item，而这恰恰呼应了我们最初定下的优化指标。在该场景下，我们学习一段用户观看视频的序列，预测对下一视频喜欢的概率。

2．深宽模型，主要是相对传统的机器学习模型而言的，如图 5-30 所示。传统的机器学习模型多为宽模型，即广义线性模型与特征海洋战术的结合。现在较为火热的深度神经网络为深模型。深宽模型即两者的结合，深模型和宽模型以及最终融合的权重放在一个模型训练流程中，不存在分阶段训练，直接对目标函数负责，端到端更加简洁。非常适合高维稀疏特征的推荐场景，发扬了稀疏特征的可解释性加上深度模型的泛化性能，双剑合璧。

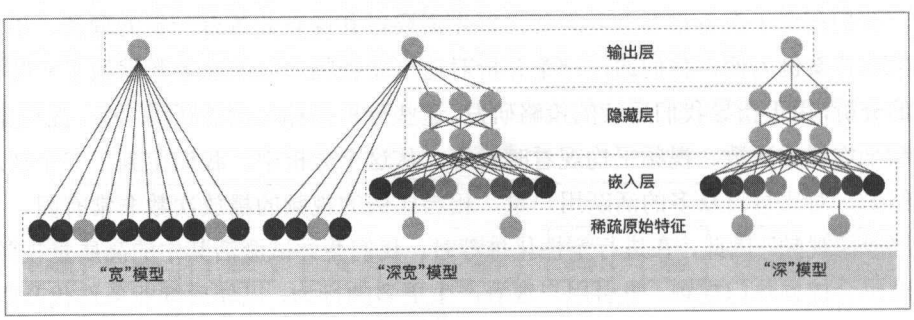


图 5-30　宽模型与深模型（图片来源于网络）

3．主题模型的采用，主要是因为隐式反馈不能解决标题党的问题。尤其在短视频行业，高亮的标题与“三俗”的图片都会吸引用户点击，对平台的长期价值是有损的。我们采用主题模型一方面可以改善推荐结果中的多样性要求，另一方面也可识别标题党对其降权。

实验与迭代

除了基于数据的情况选择合适的模型以外，实验与迭代对于一个个性化推荐

系统也是至关重要的。

秉承数据驱动的理念，我们在每一次策略上线时都会创建一组 A/B 测试，借助我们的实验分流系统，根据行为所持有的实验编号即可在神策分析平台中实时追踪、对比上文提到的几个关键的迭代和优化指标，并一次跟踪实验效果。根据实验的最终效果，最终以逐步开大流量的方式来最终完成线上策略的迭代。

第 6 章

各行业实践数据分析全过程

互联网金融数据驱动实践

在互联网金融发展的过程中，国内互联网金融呈现出多种多样的业务模式和运行机制。我国金融市场开放后，外资银行逐步进入国内市场，客户对金融行业服务和产品的需求日益多样化，因此面临着新一轮的挑战与发展机遇。金融行业日益重视用户行为分析，为了将不断增长的结构化和非结构化数据源进行整合分析，能否成功释放数据价值，实现金融业务创新，已经成为决定金融业未来发展成败的关键因素。

本节以 Formax 金融圈[1]为例，介绍其如何构建高效金融客户分析体系。Formax 金融圈是 Formax Group Limited （以下简称“金融圈”）旗下的在线金融理财交易社区，涵盖股票、基金、P2P、外汇、贵金属等金融产品，是一站式的金融服务平台。金融圈融合 SNS 社区和投资组合理念，用户不仅可以在金融圈进行投资交易，还可以交流互动，便捷地创建、出售和购买股票及外汇的投资组合，筛选出优秀投资经理和民间高手创建的组合，选择跟随购买该组合，和投资高手一样获取收益。在数据驱动的道路上，Formax 金融圈达到了以下几个目标。

1. 构建高效金融客户分析体系，实现数据驱动产品优化与科学决策。

2. 精准勾勒客户画像，多维度监控增强风控模型。

3. 搭建良性循环的精细化运营体系，精准渠道追踪，营销效果可实时观察、衡量。

---

[1] 内容因涉嫌商业机密，所涉数字均为虚拟。

实践案例

接下来，我们将详细介绍该案例的全过程。

需求梳理

科学的数据采集方式要源于企业的业务需求。金融圈公司内部按照事业部进行业务线的划分，每个事业部负责一条业务线，并由独立的运营、产品和技术人员来负责，形成统一的 APP —— 金融圈。业务部门希望可以将行为数据与业务数据进行打通，以实现更精细化的运营。金融圈 APP 针对其实际情况，梳理其主要的数据分析需求。

1．对公共平台（即金融圈 APP）用户情况精细化分析需求。

金融圈的公共平台涵盖公司各条业务线，这些业务线的需求由该平台统一整体对接。通过数据分析，希望能够评估整个平台公共功能的使用情况，包括以下几个方面。

• 用户情况：包括独立访客、页面访客、活跃用户数、新增用户数、注册用户数及各业务的活跃用户数等；

• 产品使用情况：包括平均使用时长、访问时长分布、人均访问页面数、跳出率及主要页面的 PV 等；

• 核心功能转化漏斗：包括注册流程、绑卡流程及出入金流程；

• 私信使用情况：包括不同类型人群的数量、使用未成功的数量、私信条数分布及发送私信时间分布等；

• 首页功能模块的使用情况：各个模块和功能的点击等；

• 用户关注情况：关注的类型分为自选股、牛人、栏目订阅号，不同类型的关注人数及新增人数等。

2．Life 平台精细化分析需求。

Life 平台是金融圈内部一个具有电商性质的平台，用户可以使用积分或者货币来换取商品。Life 平台的数据需求点包括以下几点。

• 用户及产品使用情况：包括 PV、UV、新增用户数、使用该功能的时长、不同页面的停留时长、跳出率等；

• 商品交易：包括各类目、各商品的浏览情况，各类目、各商品的交易情况，交易的转化漏斗，商品复购情况等；

• 订单数据：包括订单量、平均发货时间、平均送达时间等。

3. 关于外汇业务线的精细化分析需求。

外汇是金融圈 APP 上的频道之一。除了提供基本的外汇交易功能，还提供 Copymaster（金融圈外汇跟单社区，是 Formax 金融圈第一款面向全球金融交易社会化产品，汇聚了全球外汇投资高手）外汇交易工具，根据平台上真实交易收益筛选排名，普通投资者可以对特定的外汇投资高手使用“复制”功能，以期望实现最大的收益。因此在该功能中有高投资者和普通投资用户两类角色。除一些基础的产品使用情况外，伴随着交易对两类角色的分析，企业希望了解到以下情况。

• 产品使用情况：PV、UV、新增用户、注册用户数、开户用户数、入金用户数、各页面的浏览、核心功能的点击及使用时长等；

• 牛人投资者：申请牛人数、新增牛人数、放弃牛人数、在绑牛人数、牛人盈利比及牛人核心行为统计；

• 普通投资者：复制人数、成功复制人数及复制金额等；

• 交易数据：出入金人数、出入金金额、交易人数、交易金额及交易产品；

• 核心漏斗：注册开户流程、入金流程、交易流程及复制流程。

4. 对 P2P 理财业务线的精细化分析需求。

针对 P2P 理财业务线的精细化分析需求方面，我们希望关注以下几点。

• 产品使用情况：PV、UV、新增用户数、核心功能的点击情况、不同页面的停留时长及跳出率等；

• 交易情况：出入金情况、交易人数、交易金额、复投情况、优惠券使用情况、投资到期后续行为及债券转让情况等；

• 核心漏斗情况：注册转化、购买理财转化、购买债券转化及申请转让漏斗等。

5. 对股票业务线的精细化分析需求。

和外汇类似，利用 Forbag 股票组合工具，可一键购买或卖出专业投资经理或者民间高手创建的组合，需要关注以下几点。

- 推送追踪：接收 push、点击 push 链接；
- 社区资讯：点击、加载、评论、点赞、分享、分布；
- 交易：出入金情况，股票的买入、卖出、撤单、沽空；
- 跟单行为：跟单的牛人；
- 核心转化漏斗：开户流程、入金流程及出金流程等。

事件设计

根据以上的需求点，我们针对其实际业务情况和实际数据分析需求，做出了事件设计方案的建议。

1．针对公共平台（即金融圈 APP）用户情况精细化分析需求，事件设计包括启动和退出、APP 浏览页面、APP 元素点击、激活 APP、注册和登录、实名认证、绑定银行卡、入金和出金、分享等。

2．针对 Life 平台精细化分析需求，进行了浏览页、提交订单、支付订单成功、发货和收货等事件设计。

3．关于外汇业务线的精细化分析需求，进行包括外汇开户流程事件、申请外汇高投资者事件和放弃高投资者资格、购买外汇产品、购买外汇保收产品、外汇跟单等事件设计。

4．针对 P2P 理财业务线的精细化分析需求，设计了点击理财产品、提交投资信息、支付投资项目、投资成功、投资到期、领取优惠券、债权转让等事件。

5．按照股票业务线精细化分析需求，针对开户的每一步流程、浏览股票资讯和资讯的点赞、评论、分享、发布，挂单、撤单、完成交易，以及高投资者跟单事件设计等。

以上事件包含丰富的属性，我们结合用户属性，用来标记事件发生时的行为和用户特征。如外汇跟单事件中，包含高投资者类型、高投资者 ID 等属性，从而去分析不同牛人的跟单情况。再如 P2P 理财相关事件中，包含理财产品类型、理财产品名称、收益方式、投资期限、年化收益率、投资金额、优惠券 ID、优惠券类型、优惠券金额、实际支付金额、投资收益、支付方式等属性，从而可以对投资行为进行多维分析，了解不同产品类型、不同产品的投资情况，不同投资期限和收益率的投资分布，结合领取优惠券的行为去衡量优惠券的发放效果。

通过元素点击和页面浏览事件，可以采集APP中所有的元素点击和页面浏览，通过元素的内容、所在页面的名称等属性区分用户点击或浏览的是哪一个页面。这些事件作为自定义事件的补充，实现一些PV、UV、平均使用时长、平均访问深度、跳出率，各功能的点击情况等一些常规需求。

上述事件是金融圈APP前期的事件设计方案，随着该企业的业务发展、对事件设计的理解、需求的变化对事件设计又进行了优化和调整。出于对客户隐私的保护，这里只列出一个大概思路及框架。

数据接入

数据接入阶段分为两个部分，即接入方式和埋点方式。

1．数据接入方式。

在该项目中，为保证数据接入的全面性和精准性，数据接入方式包括普通的行为数据从前端采集与后端数据采集两种。

出金、入金、投资理财产品、购买外汇产品等重要事件采集从后端进行，发放优惠券这类只有后端才有记录的事件从后端采集。一些事件如提交订单等，部分属性是前端采集的，如操作系统、地理信息等，部分属性是后端采集的，如商品品牌、商品分类、商品价格等，此时由前端将它们采集到的属性传给后端，和后端采集的信息进行拼接，统一由后端发送。

2．数据埋点规范。

由于企业业务线较多，在确定了每个事件的接入方式后，企业对埋点规范进行要求，包括以下几点。

• 事件和属性名称的规范。

对每个事件，每个属性都定义好埋点的英文名称，保证各业务线、各端传入信息的一致性。为了便于区分不同的业务线，在事件前额外增加了前缀，如理财事件的前缀P2P、外汇事件的前缀forex、股票事件的前缀stock。规范化命名既方便对事件的管理，也方便后续的分析使用。

• 事件采集时机的规范。

明确好每个事件的采集时机，如元素点击事件，是该元素在前端被点击时触发，而交易成功类事件，如股票交易、购买P2P理财产品成功等，则是在服务端

返回了成功信息后才触发。精准的采集时机，使开发人员更加明确，减少了不必要的沟通成本，并且保证数据的准确性。

- 属性采集范围的规范。

当同一事件多端采集属性不一致时要明确，以浏览页面为例，该企业平台有一套适用各端标准的页面 ID 体系，他们希望能将各端的浏览页面行为进行统一分析。因此，他们没有全部采用可自动采集的 PageView 和 APPViewScreen 事件，而是在 Web 端采用 PageView，在 APP 端手动埋点。而 PageView 中有很多预置采集的属性是 APP 上没有的，对于这类属性需要明确说明，以免给开发带来困扰。

此外，特殊属性的取值范围要明确，有些事件是针对特定场景设计的，其中的属性取值是可以穷举的，也是后续需要分析的点，就需要明确列出。如产品经理需要了解某些特定页面的功能情况，而其他页面的不需要采集，此时就要明确需要采集的是哪些页面的哪些功能。

应用场景

场景 1：与工单系统结合，还原真实用户操作，高效化解客户诉求

对于金融行业而言，保障用户的每一笔资金安全与稳定是至关重要的。在金融企业内，任何与充值、提现等与钱关联的行为，一旦出现问题，会影响到用户的体验度和公司信誉，对企业发展造成很严重的负面影响。

金融圈使用工单系统进行客户服务，包括用于客户支持与帮助服务，处理与解决客户事务请求等。工单被送达目标服务台之后，主要处理流程包括：响应客户请求 → 听取客户反馈 → 反馈给技术人员 → 技术人员查询情况。

在整个过程中，客服人员做出一切判断和安排的来源，都是客户的描述，如用户进行了哪些操作，出现了哪些异常情况等。然而，从响应客户请求到处理请求，单纯依赖客服口述会导致信息不准确，延长客户服务周期，极易引发客户不满。

现在金融圈还原真实用户操作。如图 6-1 所示，通过个人行为序列能够非常方便地查看用户的具体操作行为。除此之外，还展示出每个行为事件的特定属性，如每个接口的回调结果，失败原因等。这样客服人员可以迅速发现问题，第一时间给予客户合理解释，快速解决客户问题，可视化用户行为操作，避免因用户描述含糊不清或错误，而延缓客服操作周期。并且及时定位异常情况，提升客户体验与企业公信力。

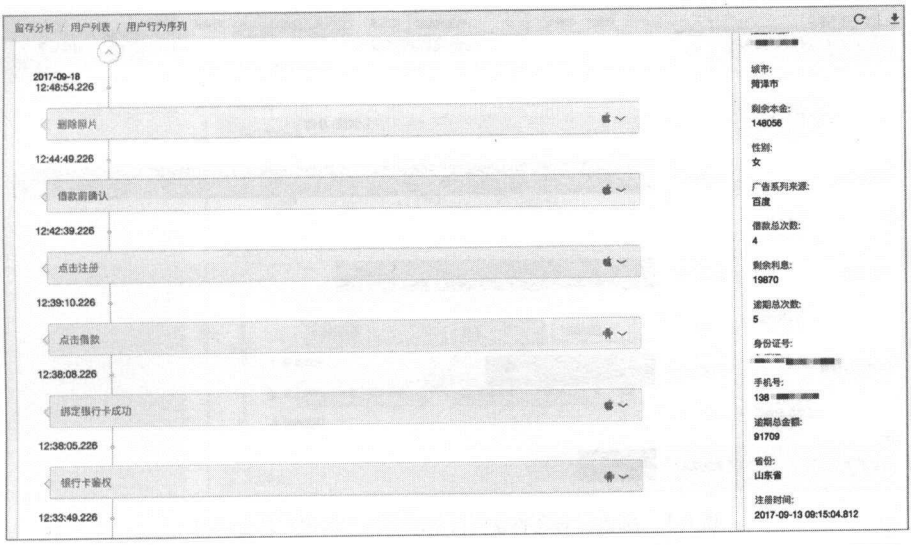


图 6-1　用户行为序列

场景 2：数据驱动定位最佳开屏主题

金融圈采集了每个页面的 ID 及该页面上所有按钮 ID 和按钮名称等相关属性，用来了解用户在 APP 上的每一步操作行为。金融圈 APP 开屏页向金融用户展示了一些营销信息或者活动信息。在 APP 运营初期，产品经理认为用户对“资金安全”的需求要远远高于用户体验。因此在开屏活动页面上会展示出“专业资质”、“多国牌照”等内容，以传递品牌安全感。

如图 6-2 所示，我们通过数据分析发现，2017 年 3 月 15 日至 3 月 30 日用户的转化率为 2.29%，结果并不理想。

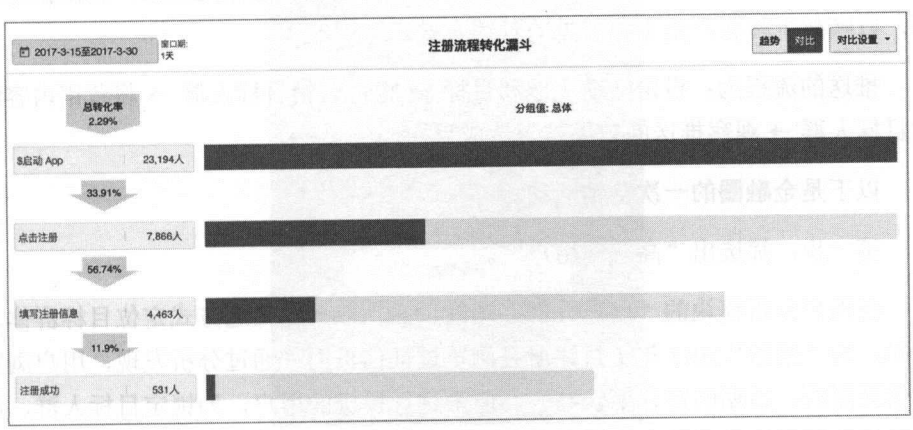


图 6-2　针对“资金安全”主题的开屏页面的转化率情况

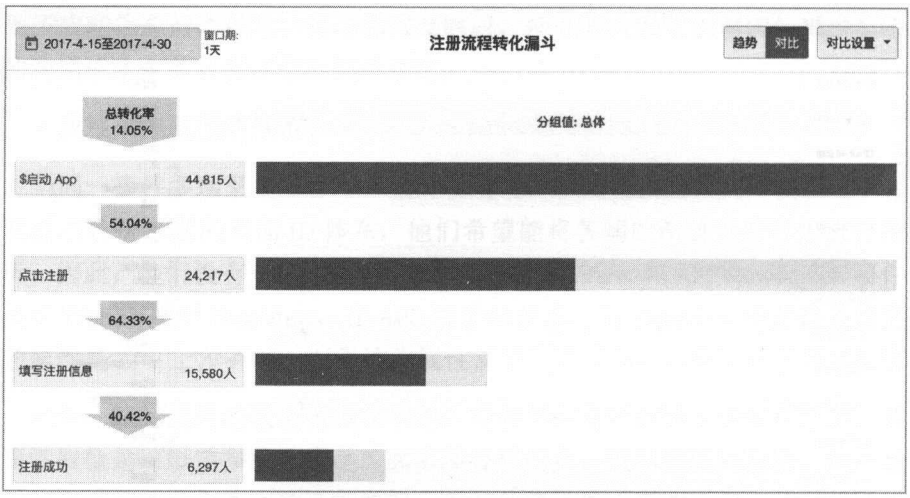


图 6-3　推出针对高收益页面活动页面后的用户转化情况

后来我们尝试推出针对以“高收益”为主题的活动页面，并经过漏斗分析发现，针对高收益的开平页面的转化率会更高，高达 14.05%，如图 6-3 所示。因此通过数据而非人员的主观判断去设计产品，这是一次较为成功的改版行为。

除此之外，依托于丰富的用户行为，我们在其他产品细节点的优化上，同样抛弃了人员主观判断的方案，依靠数据来说话，通过设计两种或多种方案，通过事件分析、漏斗分析等分析模型选择更优的方案。

场景 3：打造用户分群、精准推送、效果反馈的全流程精细化运营体系

高居不下的获客成本，增加客户黏性且延长客户的生命周期价值，是各互联网金融企业最为关心的问题，金融圈也不例外。高效、便捷地给用户精准推送内容，以唤醒沉睡客户是十分常见的营销方式。

推送的流程为：设定活动主题和目标 → 定位营销目标人群 → 将营销内容触达目标人群 → 观察推送的效果是否达成目标。

以下是金融圈的一次营销活动。

第一步，筛选出“高意向用户”。

在用户分析模块的“用户分群”功能页面，以条件筛选方式定位目标群体。例如，为“唤醒”2017 年 1 月注册且浏览过征信页面（通过分析发现，用户浏览征信页面后，后期的留存率较高），但未进行投资的用户，为锁定目标人群，可在用户分析模块的“用户分群”功能页面做如图 6-4 所示的操作。

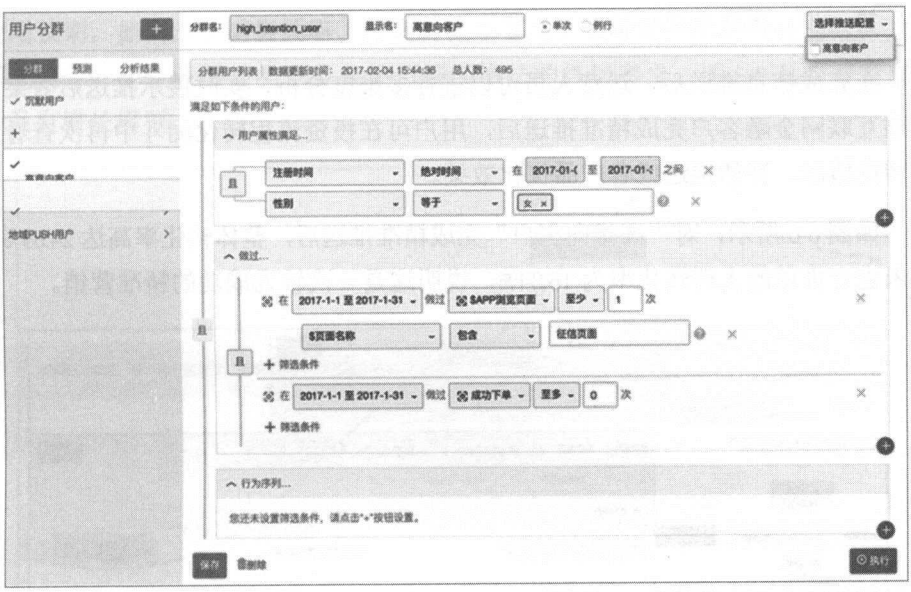


图 6-4　在“用户分群”功能页面，筛选营销目标群体

第二步，向“高意向客户”用户群体，进行信息推送。

通过用户分群功能将这部分人筛选出，可以通过短信或者站内弹窗的形式通知，并向该群体推送信息，以刺激其投资，如图 6-5 所示。

图 6-5　用户分群后，实行站内推送

第三步，推送效果评估。

在完成信息推送后，运营人员可以进行多维度分析，实时展示推送后效果。如该互联网金融客户完成精准推送后，用户可在投资流程转化漏斗中再次查看用户转化情况，评估推送或者产品优化效果。

如图 6-6 所示，对“高意向客户”完成精准推送后，整体转化率高达 24.69%，而未进行推送的人群转化率为 16.34%，说明这是一次较为成功的精准营销。

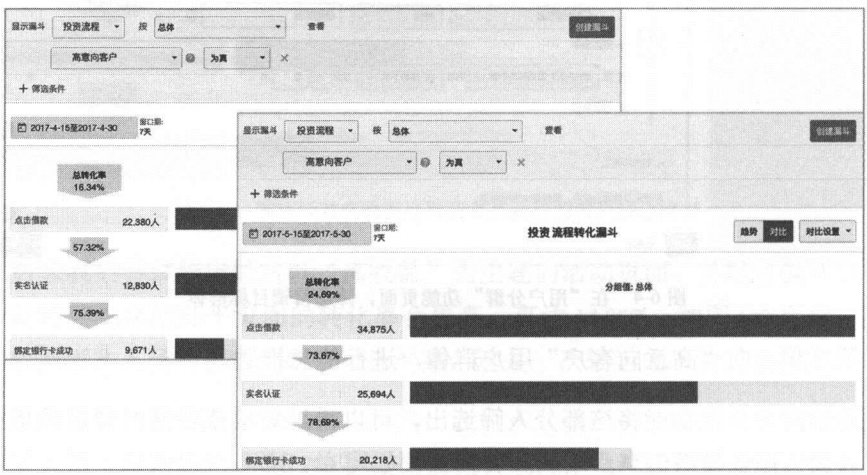


图 6-6　被推送人群与未被推送人群的总体转化率情况对比

金融圈搭建了高效、便捷、精准的营销平台。企业运营人员在可视化界面上，可依次完成多维度指标用户行为分析、用户分群、对目标人群的精准信息推送工作、实时查看推送效果的全流程精细化运营操作。

企业服务数据驱动实践

随着经济下行、人工成本上升的趋势，越来越多的企业希望通过寻找合适的企业服务来降低运营成本，同时，随着企业管理信息化日渐深入，以及传统企业拥抱互联网 + 趋势，企业服务平台将成为服务需求解决的标配。在这样的环境下，企业服务（2B）行业正在处于爆发期。

无论在企业发展的哪个阶段，提供优质的客户服务都是企业可持续增长的重中之重。企业在发展早期通常会投入大量资金来快速完成客户获取和积累，在快

速增长期，如何改善销售效率、线索转化率等成为企业管理者的关注重点，这些都成为每一个企业服务类企业业绩可持续增长的终极奥义。图 6-7 是企业发展阶段及自增长模型。

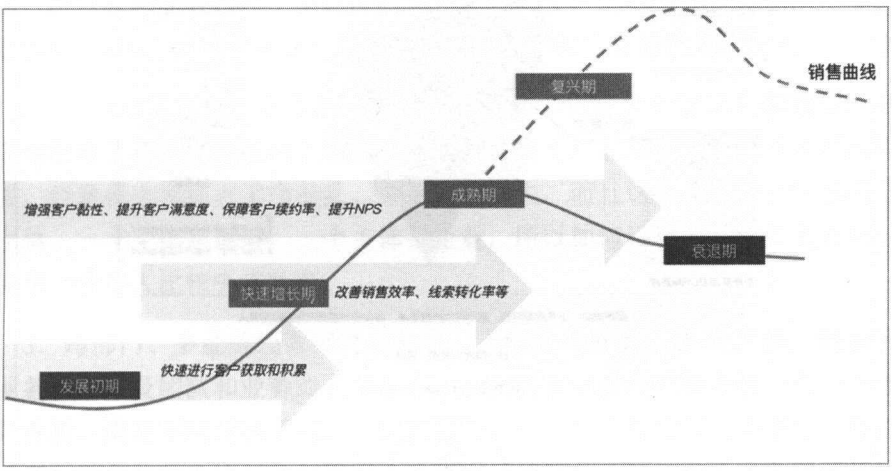


图 6-7　企业发展阶段及自然增长模型

企业服务厂商在实际转型与发展的过程中逐渐认识到，只有企业数据积累并掌控数据驱动力，才能在商业环境中占有先机。越来越多的企业面临着从“信息对称”到“数据驱动”的跨越。如果说“信息对称”能解决既有痛点，那么“数据驱动”则能挖掘并满足企业发展中的未知需求。本节将围绕数据驱动下的企业服务常见问题进行讲解。

数据驱动能够为企业服务做什么

我们知道，企业服务模式是一种全新的商业模式，不只是把软件、服务搬到网上那么简单，它意味着我们和客户之间的长期伙伴关系。企业服务的本质在于了解客户的真实业务诉求，并为其提供优质的产品与服务，帮助其走出复杂商业环境中的发展困境。深谙此本质的企业，会从客户成功角度去重构整体业务。如图 6-8 所示，我们整合历史和实时数据分析，提供更全面和深度的客户洞察。

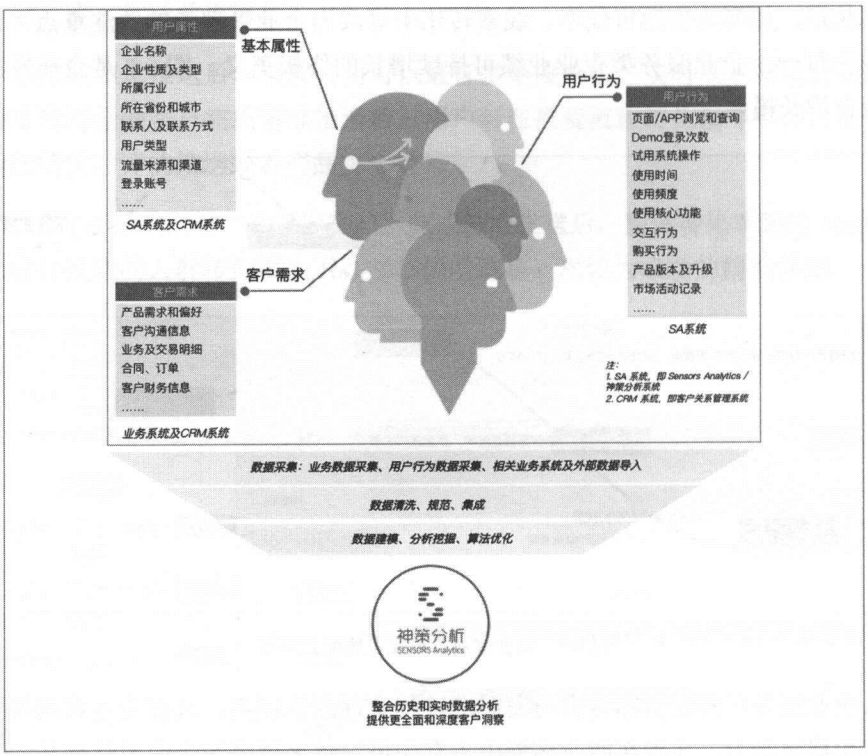


图 6-8　基于大数据的 2B 客户行为洞察

围绕以用户为中心，企业服务数据分析的核心需求有以下几点。

1．如何以较低的成本获取高质量的客户？

2．如何快速判断线索跟进优先级，有效提升销售线索转化率？

3．如何诊断易流失客群和高价值客群，实现客户全生命周期支持与管理？

4．如何根据数据提供优质的客户服务，增强客户黏性，保障客户续约率？

5．对于 B2B2C 类型的客户，如何提供给 B 端客户自有业务运营情况数据？

面临的挑战

企业服务数据驱动面临着以下挑战。

1．前端行为数据和线下、CRM、ERP 等后端业务数据无法打通。对于企业管理者而言，了解各个推广渠道，分别带来的流量和用户量是多少，每个渠道带

来的用户的后续转化和留存情况如何，是最基本的诉求。但是我们最终关心的转化不仅仅只限于注册，我们希望了解到该批线索用户有没有转化为客户，有没有最终付费。遗憾的是，大部分企业的现状是营销推广数据与企业 CRM 系统没有打通，所以就不能衡量每个渠道最终真正完成付费转化的有多少，从而不能筛选出最优质、ROI（投资回报率）最高的推广渠道，优化营销投入产出比。

2．产品功能复杂，动辄上千个埋点，不知如何定义和管理数据模型。企业服务相较电子商务、互联网金融等行业而言，由于产品功能复杂，涉及产品模块众多，经常会出现埋点无序混乱、数据采集缺失，而且这整个过程会涉及两个用户主体，一个主体是用户，一个主体是企业，所以如何设计数据模型更有利于分析就是一件令人比较头疼的事。

3．跨部门、多业务线数据完全独立，无法全局分析。2B 类产品一般会有多条业务线，涉及团队和业务线人员众多，如何将多条业务线整合统一在一个平台进行分析，满足不同团队不同人员按需提数，而不是给开发提出需求后，要排队等排期。客户的整体情况及健康度、渗透率等基础分析都依赖于多条业务线统一分析。

数据应用的阶段

企业服务的数据应用分为 4 个阶段，分别为启动阶段、黏性阶段、增长阶段和营收阶段，企业在每个阶段的关注点有较大差异。

启动阶段

该阶段主要关注市场验证，关注如何达到 PMF（Product-market fit，产品 - 市场匹配）。我们要获取一部分用户进行市场验证（这里获取的用户主要指种子用户），邀请种子用户使用企业推出的最小可行化产品，衡量最小可行化产品是否有效，最重要的数据指标与用户参与度有关。用户真的在使用这款产品吗？他们是如何使用产品的？他们使用了产品的全部功能还是部分功能？产品的使用情况和用户行为是否符合我们的预期？

黏性阶段

对企业服务行业来说，留存是根本，关注用户整个生命周期价值，更重要的是客户成功，即客户是否能够更好地使用产品，是否再续约、升级销售。

新用户的留存曲线一般分为三个阶段：震荡期、选择期和稳定期。绝大多数新用户在一开始的震荡期就流失了，在选择期部分用户找到了产品的价值，然后慢慢稳定下来。所以将核心价值功能尽可能直接地展示给新用户，增加用户使用核心功能的频率，让用户尽早发现产品价值，提升前两个阶段的留存曲线是非常重要的。要达到该目标，应该做到以下关键点。

1．找到影响新用户留存的关键功能。

通过对比不同产品功能（功能模块）的留存度，可以很容易发现产品的核心或高价值点，留存度高的产品功能的价值也会较高。通过产品的设计，引导新用户发现和使用这些核心功能模块，尽可能早地为用户创造业务价值，从而提升新用户的留存率。

2．从用户行为监控用户活跃度。

业务不会揭示问题，用户行为会揭示问题。从产品数据了解众多客户需求，分析用户到底怎么使用产品，用哪些功能，不用哪些功能，从而进行产品调整改善（融入产品开发过程中），发现新需求点进行针对性开发。企业从众多功能中梳理出核心功能，看核心功能的使用情况，以及使用频率的变化趋势。

3．提高渗透率和利用率。

产品销售出去后，确保要覆盖更多的企业员工数。同时要衡量用户的使用情况，如企信这类产品，一两个月内发1万条信息或搜索信息说明客户在积极使用，随着用户使用时间延长，会推更多服务。

团队成员密切配合教育员工如何有效使用我们产品，慢慢让更多团队成员使用，同时要避免使用混乱，因为一旦使用的人数过多，教育的成本就会增多，没被教育的用户不能很好地使用产品，间接影响口碑。

4．引导客户使用高价值功能。

产品上某个功能对用户有高价值，ROI上有提升，会促使客户更好地使用。如应用推荐，用户用得怎么样，是否可以更好地优化这个价值。

5．从客户业务的流程触发视角去使用产品。

企业需要反思，客户业务的核心业务流程是否可以用自家产品来实现，业务和产品的融合度如何，关于该问题的解答需要对整个业务流程进行拆分监控，从而形成一个核心路径转化漏斗，了解每一步的转化率和流失率，每一步流失的用

户去向何处。如果某些流程没有用到产品的话，那么我们就需要分析线上线下的行为是否存在割裂而导致不能高效运营的情况。

企业服务的业务流程很复杂，可以通过核心路径转化漏斗来监测业务流程是否发生变化。开始几周内，企业也可以更好地预测业务流程，如果发现客户不再使用该流程，要及时和客户沟通，他们到底是要解决什么问题？

增长阶段

处于该阶段的企业开始关注NPS（Net Promoter Score，净推荐值）、NSM（North star metric，北极星指标）等唯一指标，关注获取线索的数量、质量，通过追踪并分类，从而优化L2C2B[1]效率，通过反馈采取一定行动。

相较于2C的产品来讲，企业服务领域的客户获取成本是相当高的，因为获取一个B端客户需要市场、销售、技术支持等多方跟进，消耗大量人力、精力。图6-9为线索转化流程图。

在B2B行业里，购买周期通常漫长而复杂，且涉及多个阶段——需求评估、对比、评价和购买。因此，将销售人员与特定营销渠道关联起来就会非常困难。买家通常会与内容和品牌出现多次互动后，才会联系企业或者下定决心购买。了解顾客购买周期，并提供帮助他们满足各阶段需求的内容，实施正式的潜在客户培养计划，有助于定义和最大化营销活动的价值。培养计划的制定因阶段不同而异。

图6-9　线索转化流程图

[1] L2C2B，Lead-to-cash-to-behavior，从线索到现金，到行动。

1．推广、集客、引流阶段：品牌布点，获取销售线索。

该阶段会进行品牌布点，获取销售线索。通过品牌的布点，引导潜在用户关注品牌，并通过新闻媒体、软文、DSP、新媒体、线索营销等手段使用户产生兴趣。用户产生兴趣后会通过搜索引擎进行相关信息搜索，这也是大部分公司进行付费SEM推广、优化网站SEO的原因，因为在搜索引擎层面会截流一大部分流量，同时通过SEM推广、SEO优化，对特定垂直行业、地域的目标客户部门或公司来进行市场定位，实现饱和营销。

2．口碑影响阶段：精细社群运营。

用户会通过问答平台、文库、百科等途径进行信息收集、评估、决策，因此企业应该精心经营这些平台，实现流量的导流。此时也可以通过DSP、EDM进行再影响。

3．官网平台承载阶段：让销售模型零阻力。

企业官网是企业网络营销和形象宣传的平台，相当于企业的网络名片，可以起到辅助线索转化的作用。我们知道，官网网页优化重点优化激活过程，提高线索转化率。一般会经历如下漏斗转化：点击申请试用→提交试用申请→查看Demo → Demo登录。一个优质企业的企业官网，能够做到让销售模型零阻力。主要通过第一次客户成功和促进线索转化两个方面实现。

第一次客户成功是指客户刚开始和产品接触时，网上Demo要简洁、使用便利、体验良好，尽量让客户30分钟内就可以开始使用。分析免费体验用户的使用情况，不论是内部使用（个人使用）还是外部使用，是否有真实使用意愿，是否会付费，从符合产品使用场景的客户去搜寻，效果会比较好，否则就是在浪费资源。

促进线索转化是指我们要定位想获得的结果，设置转化目标，比如注册、成功提交试用申请、Demo登录等。

我们如何在降低客户获取成本的同时高效获取客户呢？

1．在投放的多个广告渠道中，找到更好的渠道并加大投入。

对于广告渠道的标记，我们推荐用业界比较通用的UTM字段来进行标记，从而保持Web、Android和iOS三个平台在分析上的一致性，各字段释义如表6-1所示。

表 6-1 渠道营销标记

中文名	英文名	含 义	示 例
广告系列来源	utm_source	标记网站、邮箱、应用等来源	utm_source=baidu
广告系列媒介	utm_medium	标记 Banner、CPC 等广告形式	utm_medium=CPC
广告系列字词	utm_term	标记广告关键词，主要用于 SEM 投放	utm_term=shoes
广告系列内容	utm_content	主要用于 A/B 测试标记同一广告间细微差别	utm_content=logolink
广告系列名称	utm_campaign	标记广告或运营活动整体的名称	utm_campaign=spring

2. 优化激活过程，提高线索转化率。

梳理线索激活转化过程，看每一步的转化和流失，降低流失率，提高转化率。假如以 Demo 登录作为转化目标的话，那么转化漏斗如下：点击申请试用→提交试用申请→查看 Demo → Demo 登录。

3. 老客户续签和向上销售。

找到新客户比维持老客户难度高 7 倍，向上销售是根据既有客户过去的消费喜好，提供更高价值的产品或服务，刺激客户做更多的消费。如果来年的客户来自老客户续签和向上销售的比例很大的话，那么也就间接降低了获客成本。

4. 病毒性传播。

在得知用户开始反复使用你的产品后，就可以着手发展用户基数了。该阶段对应海盗法则的自传播阶段，进入病毒性阶段后，即可重点关注用户获取与增长，但同时也要留意产品的黏性。在病毒式传播阶段，可以通过以下指标来衡量病毒式传播的效果。

• 病毒式传播系数

病毒式传播系数即每位现有用户能够成功转化的新用户数。

病毒式传播系数＝邀请率 × 邀请的接受率

邀请率＝发出的邀请数 / 现有用户数

邀请的接受率＝新注册数或新用户数 / 总邀请数

• 病毒传播周期

病毒传播周期是从使用产品到邀请他们加入产品所用的时间。如果从使用产品到邀请他人加入只需一天时间，增长速度就会非常快。相反，如果新用户需要几个月的时间才会邀请他人加入，增长速度就会慢很多。

• 净推荐者比例

NPS 指用户向朋友推荐你产品的可能性有多大，并比较极力推荐者和不愿推荐者的人数。该比例是病毒性的极好表现，因为它表明了有哪些客户会成为你的样板客户、推荐者或在营销宣传中现身说法。

5．采取客户分层精细化运营。

通过用户分群功能，根据客户使用产品的频率或者活跃天数对客户进行分层，包括高活跃度客户、一般活跃客户、流失风险客户，以及已经流失客户。

通过公司规模对客户分层，包括中小企业、大企业、缓慢型公司，以及快速增长型公司。

通过公司员工性质对客户分层，包括核心人员和一般人员。

通过客户性质对客户分层，包括试用客户和正式客户。

通过购买产品版本对客户分层，包括高级版、基础版、单机版及集群版。

营收 / 规模化阶段

客户终身价值和客户获取成本推动着企业的增长，同时，还会通过试验试图以更低的成本获取更多的忠诚用户，调整定价机制、收费时间及收费的服务种类。切记即便你在某个重要方面有着健康的增长速率，如用户数量或参与度，如果不能将其转化为金钱并支付成本的话，依然没有太大的意义。客户成功团队要持续跟踪客户在与公司的关系中所占据的价值，即客户终身价值。

在此阶段，应该重视以下指标。

1．收入来源分析。对收入比例做一个大致的划分。如 Salck 公司 70% 的收入来自中小企业，30% 的收入来自大企业。

2．续签。保证大部分客户在第二年能够续签，流失率控制在 5% ～ 7%。

3．向上销售和交叉销售。了解客户的使用情况，有针对性地向上销售和交叉销售。

当付费引擎状态良好，客户获取成本少于客户终身价值的 1/3 时，即可加大投入、开始扩张。低占比是付费投入回报率的积极信号。

在营收阶段，如何怎样评估此阶段公司发展情况？通常的标准是什么呢？

1．客户终身价值

客户终身价值＝（一年年服务费 × 毛利率）/ 客户年流失率 ＞ 30 万元

客户终身价值，这是一条底线。根据国内一些较成功的企业级服务公司的经验数据，客户终身价值 大于 30 万才能循环支撑起正常的获客，包括市场费用、销售成本、人力支出和有效激励。如果平均企业生命周期是 5 年，那么每年的客单价要大于 6 万，这样就大于 30 万了。

如果你花在获取新客户上的钱少于客户终身价值的 1/3 的话，那么进展就算是不错。

2．回报期（即多长时间收回获客成本）

回报期＝用户获取成本 / 客户月毛利率（毛利率需刨除云成本、营销成本等）。最好能 1 年内收回获客成本，平均时间为 18 个月，如果推迟 2 到 3 年就会有问题。

3．回收成本

回收成本＝（每季度回收的营收 / 每季度花在销售和市场上的钱）×4 ＞ 0.75。即第一季度 1 美元花到销售和营销上，下个季度获得 0.75 美元的收入，如果能够超过 1 则更好。

4．客户保有率

客户保有率＝（年度经常性收入 ＋向上销售收入－流失的收入）/ 年度经常性收入 ＞ 100%。当客户保有率 >100% 时，相当于没有流失用户。

5．年金额流失率

年金额流失率＝流失的收入 / 年度经常性收入＜ 20%。年金额流失率低于 20% 才算健康。

6．有效增长率

有效增长率＝收入增长率＋净利润增长率 ＞ 40%，利润率和收入增长率都有可能是负值，为负即亏损。

实践案例

以 A 公司——一家提供移动 CRM（Customer Relationship Management，客户关系管理）系统软件的企业服务商为例。A 公司坚持以客户为重，并一直有较强的数据驱动业绩增长意识，对客户获取、潜客线索管理和优化，以及客户服务等业务流程进行持续优化，从而实现整体经营绩效的提升。下面将介绍 A 公司如何通过数据驱动增强客户黏性、提升客户满意度，从而保障客户续约率和提升 NPS（Net Promoter Score，净推荐值）。

A 公司数据驱动实践经过了需求梳理、事件设计、数据接入等流程。

需求梳理

A 公司希望通过数据分析了解客户和真实业务诉求，对客户行为深度洞察，并为其提供优质的产品与服务，实现获客渠道优化、销售线索转化率提升，以及保障客户续约率，最终驱动企业业绩可持续增长。

可见，高效获取客户和提高线索转化率、用户对于产品的使用情况等这些需求的着眼点是用户，而对客户整体情况和健康度等的了解着眼点是企业。

针对用户的需求梳理，目的是希望实现官网潜客用户行为精细化分析，针对产品进行精细化分析。

针对企业的需求梳理，目的是希望实现多条业务线的交叉分析和对企业的精细化运营。

事件设计

我们根据企业的实际情况分别设计了以用户为主体和以企业为主体的两套不同的事件设计。

以用户为主体的事件设计包含针对官网潜客用户行为精细化分析和针对产品精细化分析。

在 CRM 模块操作事件中，它会涉及很多功能点和操作，那么如何设计数据模型才能高效分析呢？用一个功能点一个操作来设计一个事件吗？显然不行，这样会有很多事件。有没有可能设计一个事件就能包含 CRM 模块的所有功能点和所有操作呢？

我们需要先梳理 CRM 模块所有的模块、子模块和所有的操作类型，最终确定包含的属性有企业 ID 、模块名称、子模块名称、操作 ID 、业务类型等属性，这样就能通过这一个埋点事件，捕捉到用户在 CRM 模块的所有行为。

接下来介绍以企业为主体的事件设计。

A 公司一共有 5 条业务线，其中任何一个操作请求都会触发一条后端业务请求，这个过程涉及 3000 多个接口，任何一个接口都可以被调用，是否需要设计 3000 多个埋点事件呢？实际上 A 公司完全可以设计一个事件，通过属性的扩充去覆盖所有请求。

接下来梳理所有的接口，如果接口设计很规范的话，就能够按照一定的清洗规则对接口进行切分，最后将 3000 个接口数据清洗转化为一个埋点事件，它具有的属性有员工 ID、一级分类接口、二级分类接口、具体接口名、产品版本、Event_value 、FullAction 等。再结合丰富的用户属性，如企业 ID、企业名称、企业规模、企业分组、企业付费类别、企业一级行业、企业二级行业、注册时间、开通时间、代理商 ID、企业开通账号数、购买账号数、独立用户 ID 等，通过事件属性和用户属性的交叉分析，实现对企业的精细化运营。

数据接入

A 公司因为要统计在线数据，任何一个接口被调用都要被统计到，同时要保证发送的数据不重不漏，另外考虑到自己后台接口数据很规范，没有必要再耗费大量人力通过代码埋点的方式重新埋点，所以最终采用如下两种方式进行数据接入。

1. 通过 LogAgent 这种后端数据实时导入的方式接入数据。

按照上述定义好的事件设计格式，通过 LogAgent 这种后端数据实时导入的方式导入系统后，A 公司可能发现了另外一个问题，这些数据都高度聚合。那么如何定位到具体的功能或功能模块呢？通过虚拟事件功能，业务人员自主定义自己想看的功能。如只有完成了某些核心功能的企业才能算活跃的企业，那么就可以按照图 6-10 进行配置活跃企业数这个指标。

2．业务模块操作行为数据前端采集或者将数据仓库中的数据导入系统中。

此种采集方式较调用后端接口事件而言，能采集到用户一些更细粒度的操作行为数据，同时这些操作纯属于前端操作行为，不会返回给后台接口的数据。

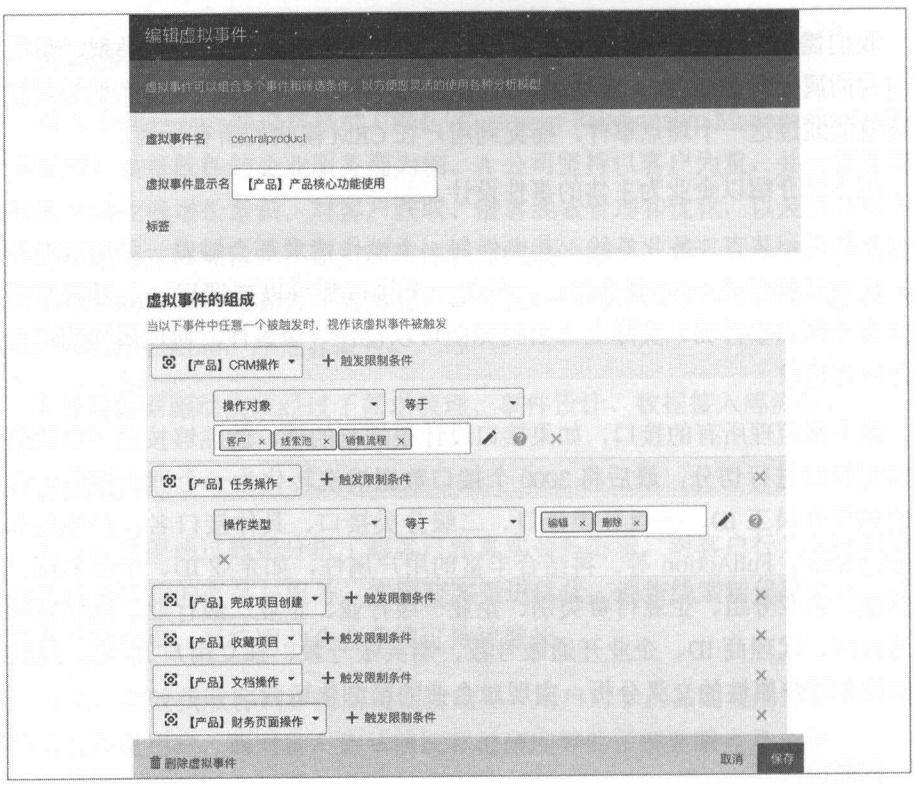


图 6-10　通过虚拟事件功能，业务人员可自主定义功能

构建企业服务的指标体系

综上需求梳理与数据采集，以及我们提及的企业服务全流程数据应用的 4 个阶段，可以看出不同创业阶段关注的核心指标有一定差异。A 企业根据发展不同阶段构建了其指标体系，该指标体系对于任何一家企业服务公司都具有参考意义。

第一阶段：启动阶段

这一阶段关注的指标包括新增注册企业数、新增开通企业数、在线企业数 / 员工数、活跃企业数 / 员工数、功能使用频率、线索转化率等。

第二阶段：黏性阶段

这一阶段关注的指标包括在线企业数 / 员工数、日活企业数 / 员工数、月活企业数 / 员工数、流失企业数、平均一个企业中有多少个用户使用、在线的员工占企业开通员工的比例、企业留存率、企业人均使用次数、功能使用频率、功能留存率等。

第三阶段：增长阶段

这一阶段关注的指标包括注册企业数、线索转化率、广告点击率、在线企业数 / 员工数、日活企业数 / 员工数、月活企业数 / 员工数、平均一个企业中有多少个用户使用、在线的员工占企业开通员工的比例、企业留存率、企业人均使用次数、渗透率等。

第四阶段：营收 / 规模化阶段

这一阶段关注的指标包括客户终身价值、平均每位客户营收 ARPU、有效增长率、免费到付费转化率、提交线索到付费转化率、回报期、回收成本、客户保有率、年金额流失率等。

值得强调的是，OMTM（One Metric That Matters，第一关键指标）在企业所处的任何阶段都无比重要，这是当前阶段的北极星指标，是高于一切、需要集中全部注意力的数字 。你可以捕捉所有的数据，但只关注其中重要的那些。对应发展所处的 4 个阶段，第一关键指标分别是活跃企业数、企业留存率、核心功能人均使用次数，以及企业质量。

应用场景

以下是 A 公司的企业数据分析的应用实践。

场景 1：一个埋点事件支撑 5 条业务线、21 个团队数据分析需求

我们知道，企业服务类企业成功的关键是促使企业用户活跃，提高企业客户的留存，降低企业客户的流失，所以 A 公司需要对企业的健康度做一个比较全面的分析，及时发现健康度不佳的企业。

A 公司一直关注活跃的企业数和员工数有多少，以及每天的变化趋势，并进行企业质量的衡量，如平均一个企业中有多少个用户使用，在线的员工占企业开通员工的比例。

这个过程存在两个难点，分别是如何定义企业在线和企业活跃。所谓在线即

产品任何一个功能点被使用则可被看做在线，问题是，既然任何一个功能被使用，企业服务产品功能相对复杂，岂不是要埋上千个事件？通过事件分析模型将上千个事件整合为一个事件再配有详细的属性就可以解决。每个业务线中团队的人员只要按照自己的需求灵活配置出自己想看的企业指标数据就可以了。

场景 2：快速判断线索跟进优先级，有效提升销售线索转化率

来自营销渠道的线索量大，CRM 系统通常记录客户基本情况，如公司名称、跟进状态、联系方式及客户所在地等，如图 6-11 所示。销售团队往往通过电话第一时间去判断客户需求、购买意愿，至于每条销售线索的处理优先级、哪些需求紧急、客户赢单的可能性大小等都较难进行快速和客观判断。因此，我们将优先判断销售线索情况的关键信息，如 SaaS 公司产品 Demo 的注册、使用等行为数据引入企业 CRM 系统，辅助销售进行快速判别。

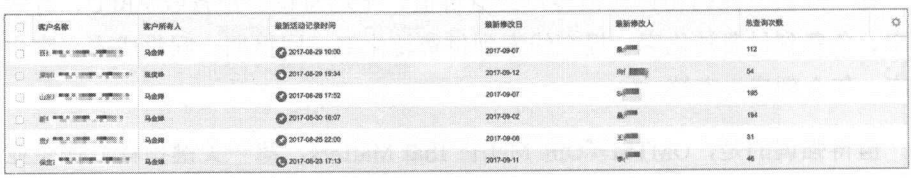


图 6-11　CRM 系统客户基本情况

我们采用后端 API 采集的方式，将用户行为数据集成到企业 CRM 系统，如图 6-12 所示。我们将采集计算好的用户行为数据，传到 CRM 上来跟踪线索，主要集成以下两个字段。

一是最近登录时间，即用户最近一次登录产品的时间。

二是总查询次数，即潜在客户在产品 Demo 上核心功能的使用查询次数。

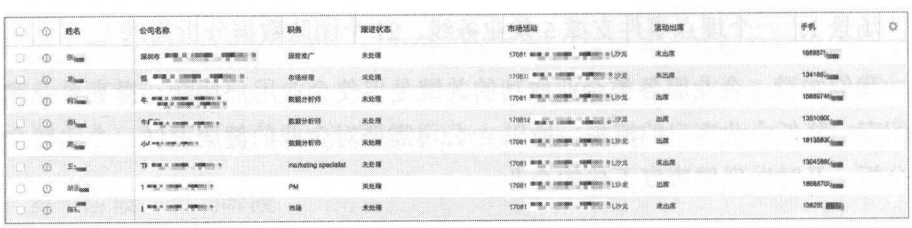


图 6-12　后端 API 集成 Demo 试用行为数据至 CRM 系统

我们通过虚拟事件定义核心功能使用次数，来计算“总查询次数”。对一款 SaaS 产品而言，其核心功能涉及的业务模块会很多，且每个业务模块下都有部分

核心功能，任意一个核心功能的使用，都可以当作客户触达了产品的价值点，所以需要通过虚拟事件的方法，将分散的各个核心功能整合为一条事件，进行整体分析。

以神策分析自身的产品为例，神策分析非常关注用户在 Demo 上“行为事件分析功能”“漏斗分析功能”“留存分析功能”“回访分析功能”“概览操作”等核心功能的使用情况，于是创建一条虚拟事件，如图 6-13 所示。创建好后，通过后端 API 采集的方式将该条事件的计算结果（总查询次数）传入 CRM，从而辅助销售团队去查看产品使用情况、快速判断用户需求和销售切入点。

图 6-13　“SA Demo 核心功能使用”事件分析

通过 CRM 和神策分析的结合，即可高效获取大量详细的销售线索相关关键信息，利用用户的行为，做到有的放矢，及时调整销售跟进策略，对不同的客户排出优先级，提高整体销售效率和有效线索转化率。

如果客户的核心功能使用次数或总查询次数从申请试用后，一直保持一个比较高的趋势的话，说明这个潜在客户转化的可能性比较高，销售团队会高优先级联系这批客户。

同时，优先选择最近登录时间比较靠前的客户，对于沉寂的客户，可以放低优先级，如果某个客户在沉寂一段时间后，某一天突然登录了，这时就可以及时跟进该客户，尽早掌握客户动态，确保最终的转化。

以的实际案例来看，销售人员拿到有价值的信息后，有针对性地跟进，在策略实施一个月后，销售线索的有效线索转化率提高了 6%，间接提高了最终的赢单率。

场景 3：提供优质的客户服务，增强客户黏性，保障客户续约率

业务不会揭示问题，用户行为会揭示问题，哪些用户是高活跃用户，哪些是高风险用户，需要从客户活跃度的角度进行监控。

A 公司通过功能细分查看不同功能的活跃度，发现大部分保持高活跃的企业用户，主要使用的功能居然是考勤签到功能，而这个功能可能是销售人员迫于绩效压力，每天例行签到的，签到之后就不使用产品了。从这可以看出定义产品活跃度的指标是不合适的，需要做出调整，只有做过核心功能的企业用户才算作活跃用户。完成核心功能的企业数和员工数的变化趋势才是客户成功团队关心的第一关键指标。如果只是浏览或点击某个功能是没有用的，只有深入使用产品的核心功能，才能发现产品价值。

我们更进一步定义活跃度，即限定每个企业至少有 3 个员工在线，并且做了核心动作中至少一个才算活跃企业。此时可以通过分布分析来看出企业活跃度分布。活跃度低的企业是重要的流失预警信号，需要重点跟进，加强培训。

场景 4：对客户分层管理，构建企业画像，实现客户全生命周期的支持与管理

A 公司灵活根据其客户使用不同产品功能的频率、活跃天数、人均使用次数等数据指标，对客户进行分层管理，详细了解每一类客户如何使用产品，然后对他们采取不同的策略。帮助企业客户成功团队和销售团队，密切关注企业状态，了解何时需要干预，实现客户全生命周期的支持与管理。

针对高活跃度客户，运营人员总结他们的使用经验，将这些经验固化下来，想办法传递给更多的客户，引导其他企业按照此方法使用，更好实现业务价值。

针对一般活跃客户，使用率一般的客户占企业用户群体大多数，运营人员常规地保持服务即可。

针对流失风险客户，运营人员首先定义一个数据模型，找到这些有流失风险的客户，再重点跟进，通过沟通、培训等方式来帮助他们。

针对已经流失客户，也就是不再使用或者不续费的客户，运营人员通过各种方式触达他们，分析流失的原因，以便于后续改进企业的产品、运营和销售。

零售行业数据驱动实践

新零售时代，一些零售翘楚正在建立起以用户为中心的业务模式，通过全量用户数据源和新兴技术来支撑全渠道业务模式的持续优化。数据采集的壁垒在打破，新零售时代线下数据发生了变化，这些都得益于更为便利的会员注册，使ID串联成为可能，此外，基于会员管理，对已有客户的挖掘，亚马逊提出的“No Lines，No Checkout”（无须排队，无须结帐）的趋势都成为可能。

新零售时代，线上线下融合要打通“三关”，具体如下。

1. 线上线下数据打通。越来越多的零售企业，如良品铺子、上海百联等企业，有线上与线下业务，线上线下数据的串联能够形成全面、完整的用户画像，这是新零售的第一步。

2. 用户行为数据与业务交易数据打通。线下POS机记录了用户的交易数据，但是交易主体情况并不知道。比如用户何时进了你的店、在哪些商品面前停留、停留了多长的时间、最后他拿起了什么商品……当这些数据不断被记录与完善后，结合交易数据就可以进行漏斗分析，了解用户整体的转化情况。

3. 全部门全场景的数据驱动。我们在强调数据驱动时，应该是企业里各个部门包括市场营销、产品运营、用户运营、管理者等都在进行数据驱动，数据驱动是全部门、全场景的事情。

接下来以O2O社区服务平台中商惠民[1]为例，介绍其如何打通线上线下数据，用户画像重塑高效流通链。中商惠民（北京）电子商务有限公司是一家以社区O2O运营服务为核心的企业，依托互联网升级社区实体店服务能力，致力于全球领先的社区O2O平台建设。他们建立了一个覆盖全国的社区电子商务服务平台和城市微物流平台，以及便民综合服务平台，是国内“互联网+”的典型代表，是互联网+社区民生的先行者。

在大数据浪潮下一直秉承数据驱动的服务理念，积极探索速度和健康的最佳平衡的发展模式。中商惠民实现线上线下数据的打通，并对数据进行高效处理和应用，掌握不断变化的商超客户实时需求，从而优化运营策略与思维，实现“人”“货”“场”的精细化管理与运营，促进服务模式向数字化转型，重塑高效供应链。

---

[1] 因涉嫌商业机密，以下场景所涉数据均为虚拟。

实践案例

以下是案例的全过程。

需求梳理

中商惠民销管中心部门、商品部门希望通过深度数据分析，为决策提供依据。“惠配通 APP”（B2B 服务体系）是中商惠民专为社区超市（便利店）经营者管理设计产品。中商惠民以新型经销者身份，对社区超市售货和进销存管理提供支持，客户遍布各类型商超。

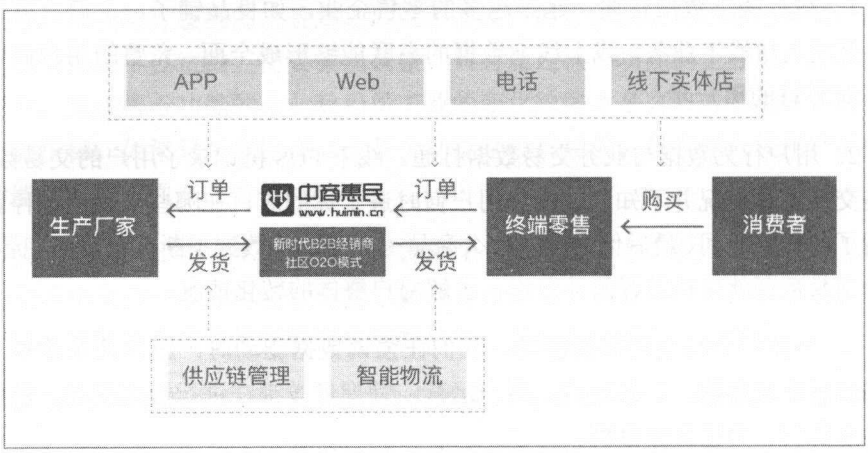


图 6-14　中商惠民 B2B 服务体系

经过多次交流，我们将需求梳理如下。

1．了解订货平台使用情况。

一直以来，中商惠民十分关注交易数据，随着业务快速发展，他们也开始重视线上业务运营状况，比如：我的商超客户是否在用 APP？是否活跃？客户应用体验如何？是否活跃？是不是每次使用都订购了货品？是否能够快速找到自己想订的货品等。

2．洞察商品销售情况。

这点也是所有零售企业的需求，零售终端（各分公司、各类型的商店、超市等）销售能力（如销售量、金额等）、刚需与高频的商品品类是企业的首要考虑因素。他们希望了解“重点客户”——满足订货量大、订货频繁、客单价高这几个条

件之一的商超客户的“刚需货”和“高频货”。

3. 评估用户活跃（付费）和留存情况。

中商惠民希望知道启动过订货 APP 的商店主，是否真的有激活（发生过首次购买），以及激活方式（自主/业代），商超客户是否都能快速找到自己想订的货品？商超客户的购买路径是否最优？商超客户的留存情况如何？是否持续在 APP 上发生订货行为？需要识别出那些即将或者已经流失的客户，再进行召回。

4. APP 各版位的使用情况。

首页推广位的效果监控是站内运营重要一环，数据的监测与分析是重要工作，它为站内优化、页面体验提升作出指导。中商惠民 APP 有众多推广位，根据主题、品类、品牌等区分，如“惠民头条”“特价专区”“套餐专区”“新品专区”等，他们非常关心这些推广位的点击率，以及后续的转化率。运营人员希望可以通过用户的点击转化率与购买转化率判断页面不同推广位置效果。

5. 业务代表的绩效考评。

业务代表是经销商的一线人员。中商惠民的业务代表需要定期拜访终端客户，了解终端客户需求，执行销售政策和促销政策，并且需要在区域经理和经销商的指导和监督下做好终端维护工作，以协助经销商完成销售指标。由于业务代表常年在外奔波，业务代表管理工作考核格外重要。为透明化管理业务代表情况，管理者需要了解我的业代人员每天在做什么？业务代表的每日行为路线？路线覆盖区域有哪些？巡店拜访签到和现场情况记录如何？等情况。

因此，中商惠民希望能够打通自己的 CRM 系统，将业务代表的行为数据也放进神策分析中进行展示，了解业务代表每日甚至是实时的业务进展和路线。同时，也希望在交易数据中，带上业务代表相关信息，将商店主的转化和付费情况作为考评业务代表的工作业绩的重点之一。

事件设计

根据以上的需求点，中商惠民进行了事件设计，在此列举部分事件设计。

1. 在用户表中记录了商店主的省份、城市、区域、店铺类型、店铺名称、店铺环境、分公司、业代 ID、首次访问时间、首次购买事件。

2. 包含 APP 使用的事件、启动和退出 APP、APP 浏览页面、APP 元素点击、

用户获取的事件、注册 & 登录、达成交易阶段的事件、加入购物车 & 收藏、提交订单 & 提交订单详情、支付订单 & 支付订单详情、取消订单 & 退货等其他关键行为事件。

3．针对 Banner 位分析，包括发放、领取、兑换和使用优惠券、充值、业务代表在 CRM 系统中的操作等。

以上事件包含丰富的属性，我们结合用户属性，标记事件发生时的行为和用户特征，如：订单相关的事件中包含订单金额、是否使用优惠券、下单入口等属性，从而可以查看各分公司（店铺）各时段商店主下单的次数、金额，使用优惠券的张数、补贴金额，再按照金额和次数的分布将商店主进行分组。

订单详情相关的数据包含每种商品的 ID、名称、品牌、类别、价格、是否折扣商品、是否套餐等属性，从而可以下钻每个订单，了解不同（类、品牌）的商品在各分公司、各地区、各店铺订了多少。

通过元素点击和页面浏览事件，我们可以采集 APP 中所有的元素点击和页面浏览，通过元素的内容、所在页面的名称等属性区分用户点击 / 浏览的是哪一个元素 / 页面。

这些事件，作为自定义事件的补充，采集的现阶段需求相对比较简单，企业只需要了解概况的行为。

数据接入

中商惠民的接入方式采用以下原则。

1．普通的行为数据从前端采集。

2．支付、充值等重要事件从后端采集。

3．发放代金券这类只有后端才有记录的事件从后端采集。

因为支付相关的事件如支付 / 提交订单、支付 / 提交订单详情等事件中，本身包含很多属性，如商品类别、商品品牌、业代名称、激活方式，这些需要在后端才能取到，因此，将所需前端能采集到的属性传给后端，和后端采集的信息进行拼接，统一由后端发送。这样既保证了数据信息的全面性，也提升了数据的精准性。

应用场景

下面，我们分别从“货”“场”和“人”几个角度进行阐述。

1．管理“货”：有的放矢，科学布局智能供应链。

库存是一场企业要赢的生存战，降低库存成本和提高企业市场反应能力是企业的目的之一。作为全球领先的O2O社区服务平台，中商惠民有丰富的上游商品资源，具备在供应链上进行深度尝试与整合的能力。

场景1：如何找到重点客户的“高频货”

这是零售企业亟待破解的难题。零售终端（商店、超市等）销售能力、刚需与高频的商品品类是企业的首要考虑因素。我们先假设满足订货量大、订货频繁、客单价高的商超客户为“重点客户”。为了解重点客户“刚需货”和“高频货”，中商惠民运营人员从商品品牌、商品类别、商品金额等分析维度出发。

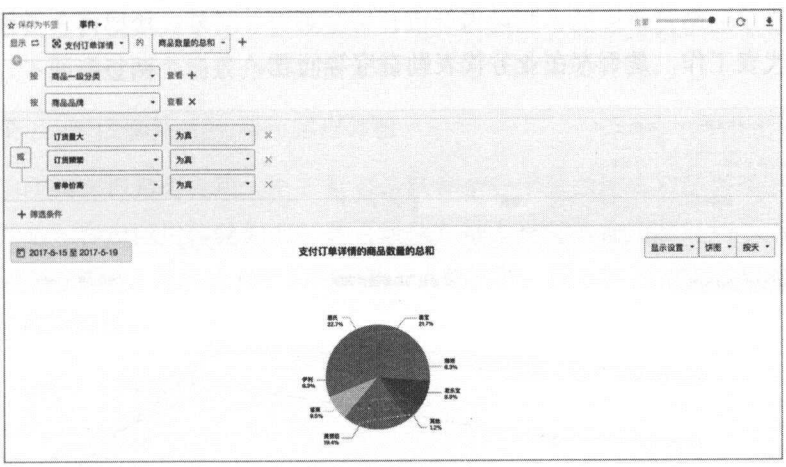


图6-15　重点客户不同乳制品品牌的商品销售

图6-15呈现出了中商惠民重点客户不同乳制品品牌的商品销售数量，这些数据表明不同品牌饮料的紧俏程度。显然惠氏、喜宝、美赞臣在乳饮品牌中属于重点客户的“高频”商品。当然，还可以从商品类别如百货、母婴等分类、商品名称如花生牛奶、文具用品、面包等判断不同类型客户的青睐。

企业同样可以掌握“非重点客户”的品牌青睐，除此之外，还可以在神策分析平台上了解这些“高频货”和“刚需货”都卖给了哪些商超。当企业掌握了不同规模的商超客户的经营品类、采购频率、销售总量、偏好分析等数据后，中商

惠民可精准勾勒用户画像，并结合实际合作品牌对商超客户进行选购引导，进而评估商品品牌铺货的合理性，如是否将特定商品放置合适的地点？铺货量是否合理？以评估供应链状况科学性。

场景 2：作为经销商，能否向品牌商交付优质业绩单

作为上游生产厂商的经销商单位，为确保以后能与品牌厂商持续稳定地合作，经销商要为合同（协议）履行期间交付的销售业绩负责。为保证阶段性的销售任务达标，经销商要实时了解各城市、各阶段的铺货情况，以实现市场渠道终端品牌销售的精细化管理。管理者可以实时了解各个商超的铺货情况，对销售预测提供一定的科学根据。

“惠氏”不仅是高频商品，而是一直以来重点铺货的乳制品品牌。图 6-16 清晰展示了各分公司对“惠氏”品牌铺货情况，可见深圳重点客户的铺货量一直处于最低点。为进一步找到原因，企业可以深入剖析深圳市场重点客户的惠氏品牌过往销售量，结合该区域乳制品偏好进行铺货调整。另外，还应深入考核深圳口的业务代表工作，是否存在业务代表勤奋度等问题。

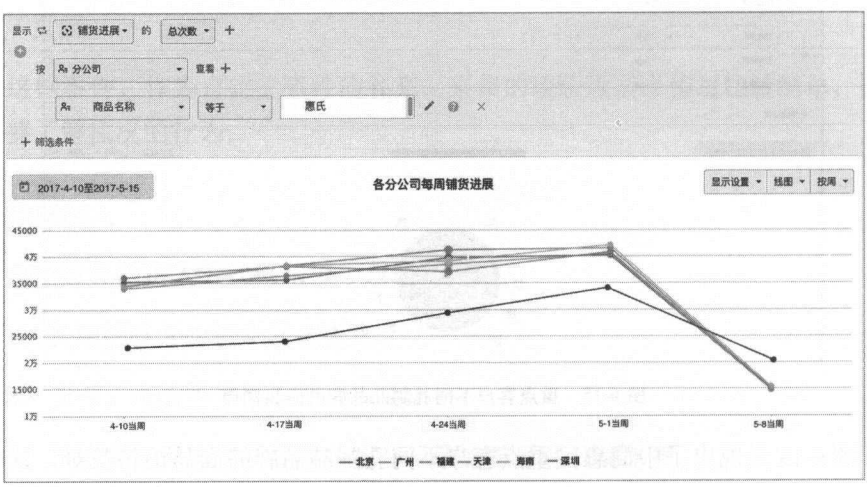


图 6-16　企业各分公司对特定商品 —— 惠氏品牌的铺货情况

综上所述，通过数据分析，中商惠民能够实时了解零售终端的销售情况，科学衡量供应链的健康度，并根据市场需求快速做出反应，及时安排采购部门提前完成物料采购、高效备产、物流配送等。同时，将生产厂商和零售终端紧密联系在一起，使消费需求数据、信息迅捷地传达给生产者和品牌商。在一定程度上，帮助生产厂商向市场需求拉动生产转变。

2．管理“场”：线上线下融合，构建多渠道互动体验。

新零售的核心是线上线下的融合。多数零售企业会通过线上电商销售以及线下实体销售，来构建线上线下多终端全零售服务场景，线上与线下体验融合与精细化运营，给客户建立多渠道的立体式互动体验。

“陈列”与“氛围”是卖场的重要衡量因素——线上线下店铺都需要考虑陈列关注商品是否易浏览、易购买、展现形式是否刺激消费欲望，其中线下店铺强调“现场管理”，包括陈列、氛围、卫生等，啤酒与尿布的故事是线下卖场布局的智慧经典所在；线上卖场（APP 或 Web 端）即需要关注页面布局是否具有购物（支付）引导性，配色与风格调性是否传递了轻松的视觉感官，用户的选购体验操作是否舒适等。

科学的数据分析可以无限逼近客户真实意愿，零售企业通过改进购买决策路径、优化列表页的体验、提升首页流量分配效率、最终提升用户的转化率。其中常用的数据分析模型为事件分析模型、分布分析模型、漏斗分析模型、点击分析模型等。下面通过两个场景介绍数据分析在“场”的管理价值。

场景 1：深度感知 APP 商超客户体验

我的商超客户是否在用 APP？是否活跃？企业十分关注客户是否热衷于使用 APP / 网站？客户应用体验如何？是否活跃？是不是每次使用都订购了货品？图 6-17 显示了 2017 年 3 月启动 APP 的用户数及人均次数，图 6-18 显示了 2017 年 3 月客户 APP 的应用时长。

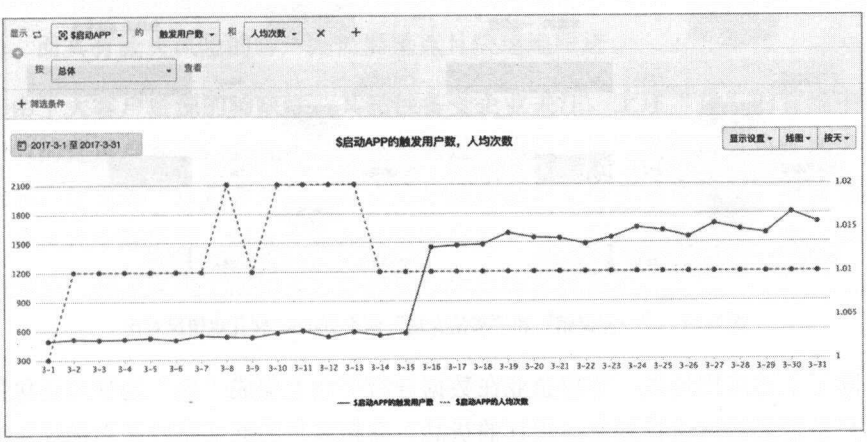


图 6-17　2017 年 3 月启动 APP 的用户数及人均次数

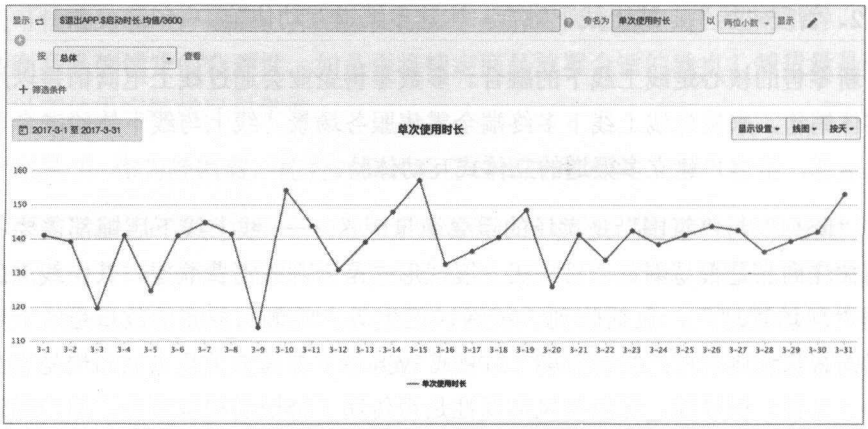


图 6-18　2017 年 3 月客户 APP 的应用时长

场景 2：科学评估站内推广位的效果

首页推广位的效果监控是站内运营重要一环，数据的监测与分析是重要工作，它为站内优化、页面体验提升作出指导。运营人员可以通过用户的点击转化率与购买转化率判断页面不同推广位置效果。图 6-19 是中商惠民首页推广位“一元促销”和“清洁专场”两个 Banner 转化率情况对比。

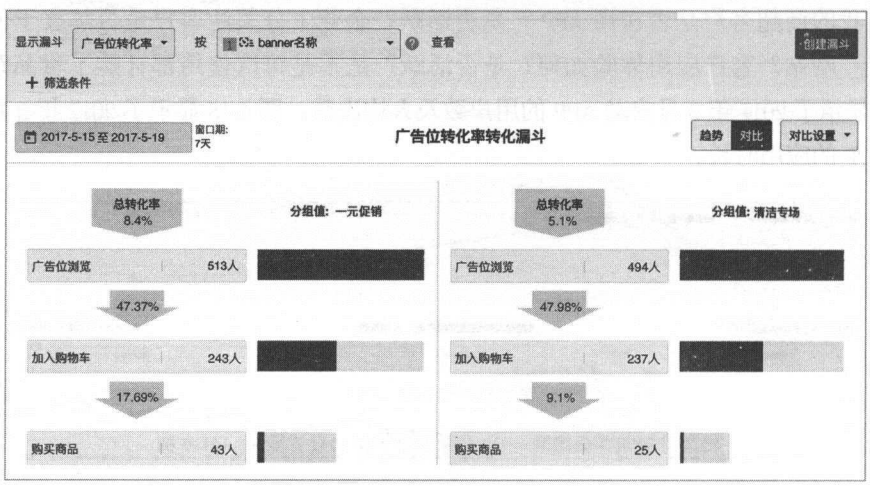


图 6-19　“一元促销”和“清洁专场”两个 Banner 转化率情况对比

除了上述应用场景，零售企业在数据分析平台上完成“场”的管理还包括商超客户是否都能快速找到自己想订的货品？商超客户的购买路径是否最优？商超客户订货意愿低的症结在哪等。

综上所述，通过数据分析，我们实现对零售线上用户的活跃度、留存率、付费转化率、商品推广效果、渠道 ROI 等深度分析。对“场”的管理价值主要包括以下几点。

• 科学评估促销专题活动效果，提升客户量、活跃度、留存率，提升商品推广策略的科学性。

• 深度感知用户体验，精细化评估网站或者 APP 运营的合理性，帮助企业科学制定运营方案。

• 了解客户产生支付行为的主路径和次路径，以及影响转化的主要因素和次要因素，有的放矢优化商超的购买路径。

3. “人”的管理：数字解密，用户画像与业务代表评估从模糊到精准。

广义上来说，“人”的管理主要包括对客户管理和业务人员管理两方面。

客户管理的重点是关注用户整个生命周期价值，更重要的是客户成功，即客户是否在更好地使用产品，是否再续约、升级销售。运营人员可以对客户分群管理，从而采取不同的策略。

用户分群分析模型能够帮助企业甄选出具有一致属性或特征的用户群体，并深度观察其行为特征。众多零售企业在神策分析平台上借助用户分群功能配合其他分析模型，能够了解到客户使用产品的频率、活跃天数、使用深度、采购趋势等数据指标，快速甄选出高活跃度客户、一般活跃客户、流失风险客户。对于处于中间环节的供应商来说，高活跃度客户成功经验能够传递给企业许多优质运营经验，而具有流失风险的客户则需要重点且快速地跟进。

由于大客户资源的稀缺性，其黏性备受企业关注。上述“货”的管理中已介绍如何筛选重点客户，此处不再赘述。通过“用户分群”功能可以将筛选出来的这批客户定义为“重点客户”。重点客户流失前兆是购买频率降低、充值金额降低、登录频率降低等，根据企业业务情况有所差异。图 6-20 是通过“用户属性”分析模型，筛选出距上次购买已经超过一个月的重点客户。

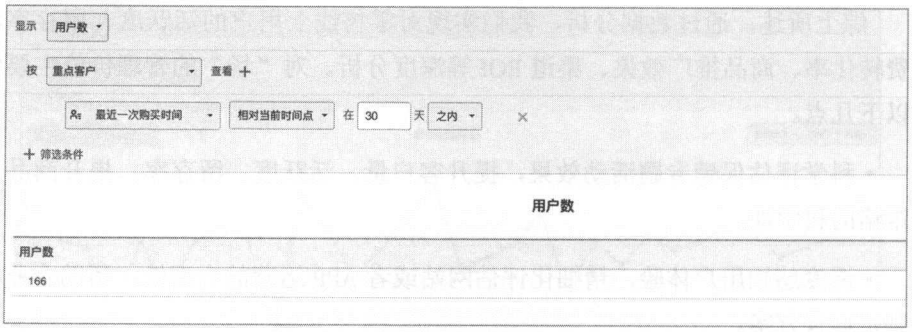


图 6-20　准流失客户群预警

图 6-21 显示有超过 166 个商超客户一个月未订货。点击数字 166，了解 166 家重点客户明细。30 天未发生购买的原因很多，有的是重点客户未流失，只是不再用 APP 下单，有的是重点客户流失了等等。此时就需要业务代表进行召回动作，无论属于哪种情况，运营人员都可以在通过个别用户行为（重点客户）序列，分别了解重点客户路径，找到重点订单量骤降的原因。

业务代表是经销商的一线人员。中商惠民的业务代表需要定期拜访终端客户，了解终端客户需求，并执行销售政策和促销政策，且需要在区域经理和经销商的指导和监督下做好终端维护工作，以协助经销商完成销售指标。由于业务代表常年在外奔波，业务代表管理工作考核格外重要。为透明化管理业务代表情况，管理者需要了解业代人员每天在做什么、业务代表的每日行为路线、路线覆盖区域有哪些、巡店拜访签到和现场情况记录如何等，如图 6-21 所示。

数据分析平台实现与 CRM 系统对接，除了高质量线索的渠道评估、把握客户需求赋予销售精准洞察力外，还实现了对业务人员的管理。[1]

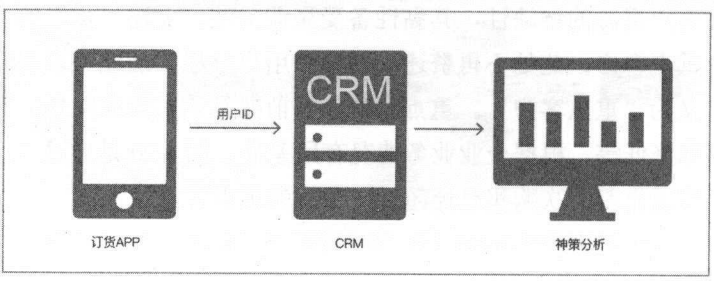


图 6-21　神策分析与 CRM 实现对接

[1] 参考文章：《神策分析和 CRM 系统结合帮助客户打造成功模式》http://www.prnasia.com/story/176606-1.shtml

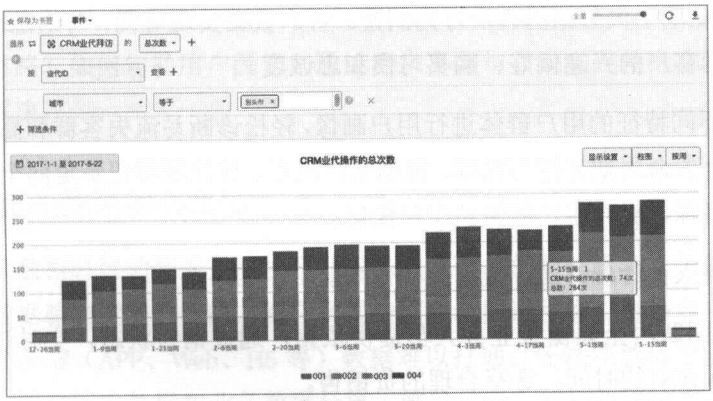


图 6-22　从 2017 年初至今，包头市 4 名业务代表的拜访客户次数情况概览

图 6-22 为 2017 年初至今，包头市 4 名业务代表的拜访客户次数概览。不难看出，单从拜访次数上说，003 的拜访勤奋度明显高于 002。不同业务代表所负责客户完成支付订单的情况也是评估业务代表绩效的重要因素，如图 6-23 所示。

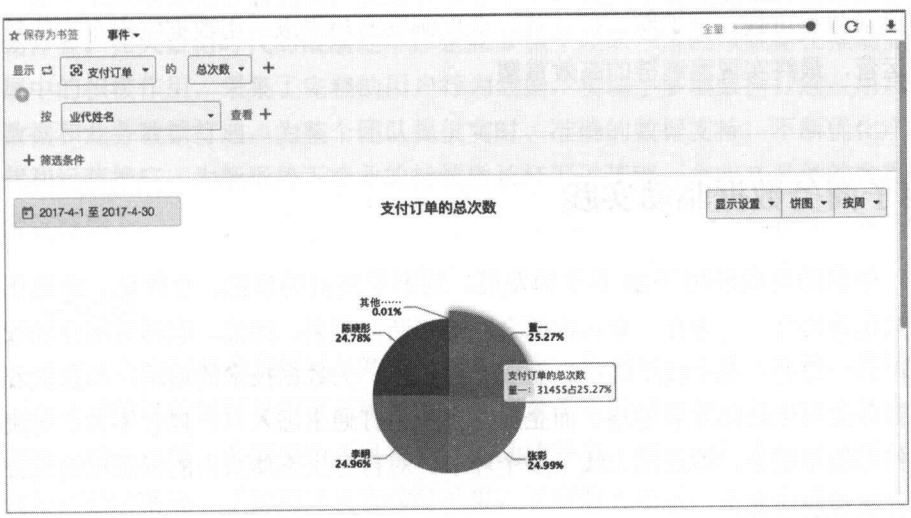


图 6-23　不同业务代表所负责客户的支付订单次数

总之，为保证业务代表考核的科学性，我们应该多维度全面地考量。神策分析与 CRM 对接后，管理者可以按时间周期汇总业务代表拜访客户情况、分布、订货情况等，并能针对性分析单个业务代表的行为线路（时间、路线内外的操作情况——拜访、上传等操作）等。

在客户管理方面，数据分析带来的价值如下。

1．建立基于大数据的客户行为洞察。客户数据实现可视、可追踪、可优化，细粒度勾勒客户的兴趣偏好、购买习惯和忠诚度。

2.为不同特征的用户群体进行用户画像，轻松诊断易流失客群和高价值客群，实现客户全生命周期支持与管理。根据客户地点、过往消费记录定向个性化推送，实现客户追踪，提升客户体验和库存优化。

在业务人员管理方面，数据分析带来的价值如下。

1．为团队内全面衡量业务人员提供依据；通过销售状态，实时进行干预，保证对的人在对的时间，选择合理的货销售。

2．企业 CRM 系统整合打通，形成从客户获取、销售成单到客户服务的完整闭环，成为企业业绩可持续增长的驱动引擎。

新零售时代，是以客户为核心的全域洞察时代。在新服务业态下，客户数据不完整，粒度粗糙等已成为零售企业发展最大羁绊。类似中商惠民的零售企业通过追求数据可视、可追踪、可优化，深度洞察商超需求，用数据实现全方位精细化运营，最终实现流通链的高效重塑。

电子商务数据驱动实践

中国的电商经历了 20 多年的发展，逐渐呈现出场景化、个性化、国际化、社交化等趋势。一方面，电商由综合网购不断向母婴、跨境、农村等细分领域发展；另一方面，线上线下结合、企业合纵连横、大数据技术的运用，都象征着电子商务走向生态化发展道路。而企业需要不断打通生态入口、耦合零售、物流、支付等场景服务，涉及线上线下多个环节，对自身生态体系内的资源重新整合，来打破行业边界。

打破企业发展经营困局：从粗放式到精细化

近 10 年来，电商行业竞争日趋激烈，渐渐从单纯依靠激情和人海战术的粗放式发展逐渐回归理性，越来越多的企业先驱者意识到要逐步提高运营效率和管理能力，实现精细化运营是企业生存和发展都面临的严峻挑战，例如市场营销的 ROI 如何进行有效评估，从而优化整体营销资源，构建营销核心竞争力？如何让大数

据连接企业内外部，提升企业生产力和经营效率？如何树立以用户为中心的思维，转变运营思路？如何勾勒用户画像，通过用户喜好与需求实现精准推荐？这些问题都需要解决。

电商企业数据驱动瓶颈

随着大数据时代的到来，很多企业纷纷建立自己的数据平台，以期实现精细化运营、数据驱动发展的目的。而在实现数据驱动的过程中，电商企业面临重重困境，例如多端（APP、Web、H5 等）数据难以打通，无法在整体平台上统一查看与分析。用户行为数据和业务数据分离，难以打通，行为数据缺乏业务属性。决策者缺乏实时、准确的数据作为决策依据。

实践案例

接下来以电商企业 B 为例，企业 B 成立至今，致力于创造简单、值得信赖的购物体验。在大数据时代的浪潮之下，企业 B 整个团队一直注重数据在推动业务发展中的积极作用，采集了完整的用户行为数据，实现了多端数据打通，用户行为数据和业务数据打通，为整个团队提供实时、准确的数据支持，不断优化产品迭代和运营推广，为新形势下企业的持续增长打下了基础。企业 B 具体的数据驱动全过程如下。

需求梳理

企业 B 各部门都希望通过从数据中挖掘信息，指导自身业务的发展。总的来说，市场营销部门需要寻找更高质量的渠道流量；运营部门希望对每一次活动了若指掌，通过对每一个细节的改进来提高活动的效果；产品部门希望通过观察用户使用产品的情况，了解用户真实体验结果，了解服务质量，指导产品每一次迭代的改进。经过反复交流沟通，我们将其核心需求梳理成以下几点。

1. 掌握各渠道引流能力，评估渠道质量。

随着互联网流量红利的消失，流量变得越来越贵，如何明确市场推广目标，如何制定分阶段目标，选择什么样的渠道，每个渠道的预算显得至关重要。在需求沟通中，企业 B 十分关注不同渠道带来的流量分别有多少？不同渠道的流量注册转化率如何？不同渠道的流量之后的购买转化率如何？不同渠道带来用户的留

存如何？不同关键词之间的流量分别有多少？不同的落地页的跳出率如何等等。

2．洞悉用户购物体验。

在用户的购物过程中，一个影响购买转化的重要因素就是购物体验。而如何提高用户的购买体验，直接关系到产品的发展，是产品开拓市场的前提和基础。企业 B 非常重视这一块数据内容，希望能从产品细节处发现可优化的点，具体表现在关注产品各个不同的购买路径转化率如何？购买路径中主要的流失点在哪里？产品服务质量如何？用户的反馈是否得到及时的回复？等等。

3．评估站内运营位效果。

站内运营位效果管理是运营管理的重要内容，同时也是运营团队进行资源评估、效果优化和内容审核的重要参考依据。在此之前，企业 B 主要通过业务数据来支持运营效果评估。在接入行为数据后，他们希望了解到不同运营位的流量点击对比，不同运营位流量的后续转化率，访问商品详情页时长，是否购买等；通过了解不同运营位效果，判断不同位置在今后的运营过程中如何进行资源分配。

4．洞察商品销售情况。

企业 B 所有产品、运营的改进最终还是要回到商品的销售情况上来，哪些品类商品销售得更好？什么样的商品页面是购买者经常访问的？用户访问次数最多的商品页面是哪些？这些商品最终是否被购买了？哪些品类、品牌或商品是用户决策时间段购买转化高的？这些问题也是在需求沟通中企业 B 提出的疑问。

事件设计

从用户的分析需求出发，企业 B 进行了针对性的事件设计，下面是部分事件设计的介绍。

1．针对市场推广的需求，事件包含 APP 启动、注册、登录、APP 激活、退出 APP、Web 端页面浏览事件，以及对应的渠道推广相关属性。

2．针对用户购物体验的需求，事件包含 APP 浏览页面、Web 浏览页面、搜索事件、浏览商品详情页、添加购物车、移出购物车、提交订单、提交订单详情、支付订单、支付订单详情、取消订单、订单发货、订单收货等购买行为路径中的关键事件等。

3．针对站内运营位效果分析，设计的事件包含各运营位模块内容的曝光 &

点击事件，优惠券的领取 & 激活、活动兑换、直播营销的播放等事件，同时在购物流程关键事件中都带上运营位模块来源信息，可以直接监控各运营位带来的效果和最终购买的转化情况。

4. 针对产品服务质量，设计了商品评价、物流评价、收藏商品、联系客服、问题反馈、产品评分等事件，以及评估服务质量的如评分、反馈等待时长等属性。

5. 用户属性方面，记录了用户的昵称、性别、省份、城市、地区、注册事件、VIP 等级等基本信息之外，还设计了用户首次访问来源、首次访问时间、最近一次访问时间、最近一次下单时间以及用户标签等信息。

除了以上这些事件本身的采集，还采集事件发生时的其他维度信息，如触发浏览商品事件时，就需要采集商品 ID、商品名称、商品款式、商品价格等，这些维度都反映了用户潜在的购买选择倾向。同时，SDK 自动采集了许多维度信息包含地区信息、是否使用 WiFi、操作系统、操作系统版本、设备型号、设备制造商等信息。结合这些多维度的信息，我们能区分不同地区的访问量、订单量，能跟踪不同地区用户的购买习惯，也能在 APP 出现异常时，结合操作系统、设备型号等信息发现定位问题机型，还能通过商品价格分布区分不同用户群体，并对其进行针对性运营活动。

通过以上事件和属性数据全面丰富的采集，企业 B 实现了行为和业务数据的整合，从而为后续在一个界面上，实现快速完整的统计和分析，打下了坚实的基础。

数据接入

在接入数据时，企业 B 遵循了以下几个基本原则。

1. 能在服务端获取的数据，在服务端采集。

2. 页面浏览、点击动作等客户端交互行为在客户端采集。

3. APP 和 H5 混合开发的内容之间互相打通。

4. 服务端采集数据，客户端传递公共属性值给服务端。

为了减少客户端埋点在网络传输过程中可能造成的异常，大部分数据采集在服务端采集，并且使用了本地备份后工具导入的方式，避免了因网络等原因导致数据上报异常问题。同时服务端采集数据的优势在于不受客户端发布新版本的影

响，事件采集中的属性增加，错误修改等可以调整。在常见的客户端采集数据的过程中，若发现数据采集异常需要等到下次发版时才能修正，且未更新新版本的用户数据将长期无法修正错误。

在常见第三方统计分析工具中，很多均采用 iOS、Android、H5 等多端数据分离采集、存储、分析。企业 B 在使用神策数据采集过程中，除了移动端、PC 端数据统一采集、存储、分析外，还使用 SDK 中的 APP 和 H5 之间用户打通功能，使混合开发的 APP 在用户匿名登录时，采用统一的用户 ID，将用户的 APP 端、H5 端的匿名行为和登录后行为连接起来，采集更完整的用户行为路径，在提升数据准确性的同时，也提升了数据的价值。

应用场景

下面是企业 B 的实际应用场景。

场景 1：破除虚假繁荣，数据驱动拉新

拉新一直是众多活动运营的目标。拉新活动方式有很多，包括“以老带新”客户口碑式的拉新、锁定目标式的地推拉新、宣传合作的线上渠道拉新等。精心策划的线上线下活动，可以快速实现拉新。

企业 B 在向海外业务扩展的过程中，策划了一次长达一个月的“A 计划”拉新活动。活动上线后第一天效果显著：日新增会员数量增加高达 100% ～ 200%，且活动当日平台的成交额提升了 20% ～ 30%。两个指标的飙升让项目参与者格外兴奋。

在数据分析领域，“总注册数”和“新增注册数”本身是一个虚荣指标，该指标随着活动力度、形式等呈现短期暴增，它能告诉你活动传递并影响了多少“新用户”，这些新用户知道你在做什么，而并不意味你的产品一定对他们有价值，还需要结合新用户的留存、转化等情况综合考量。随着活动的持续进展，配合数据分析平台的深度观察，看似美丽的外表暗藏着不少问题。

最为典型的问题是转化率较低，付费意愿较差。一周后数据统计，与自然流量相比，新注册会员的留存率与转化率均低于其 50%，从注册到浏览商品详情的转化率低于其 60%，另外，人均支付单数降低到原单数的 1/3，人均消费金额仅占自然流量的 25%。图 6-24 是活动带来的新会员与自然流量会员的转化情况。

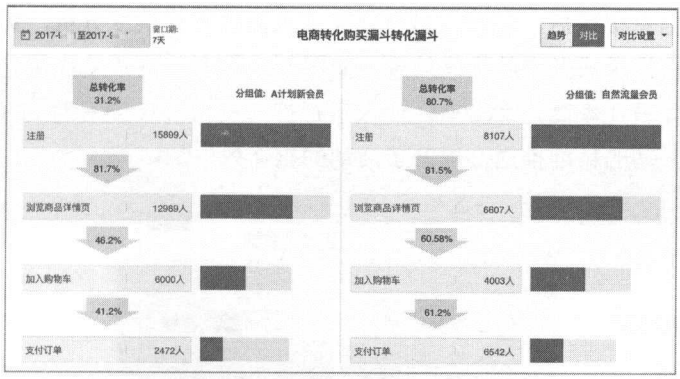


图 6-24　一周后新会员总转化率均低于自然流量 50%

这些暴露的问题体现了活动的不完善之处，拉新活动急需解决这些问题，新人的快速流失可能有很多原因，说明平台对于活动进入的新会员的吸引度不够，或者会员在产品中未能及时获取最为关注的内容。针对新用户留存、转化率低问题，活动人员进行了以下改进。

1．增加新人频道，投其所好促进用户转化。

为了增加新会员黏性，针对新会员增加一个新人频道，以店铺打折信息、精品推荐等形式针对新用户推出一系列活动。我们发现该活动对新用户的转化和留存存在很好的效果，新用户次日留存提升近 40%。

2．精准推送，用户分群促进会员留存、转化。

活动人员筛选出注册后一周内未交易的会员，这些会员是潜在的准流失会员。选择对该目标人群进行一次短信与站内的推送。图 6-25 展示了 7 天内未支付订单的新会员情况。

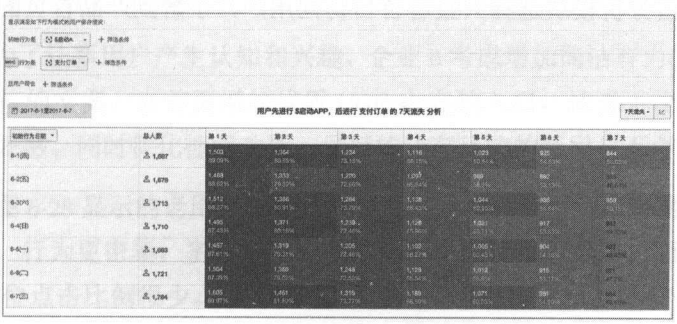


图 6-25　了解 7 天内未交易的新会员情况

在完成信息推送后，运营人员可以在主页面进行多维度分析，实时展示推送后效果，评估推送或者产品优化效果。如图 6-26 可见，对“流失用户”完成精准推送后，整体转化率高达 24.69%，而未进行推送的人群转化率为 16.34%，说明这是一次较为成功的精准推送，提升了活动的整体效果。

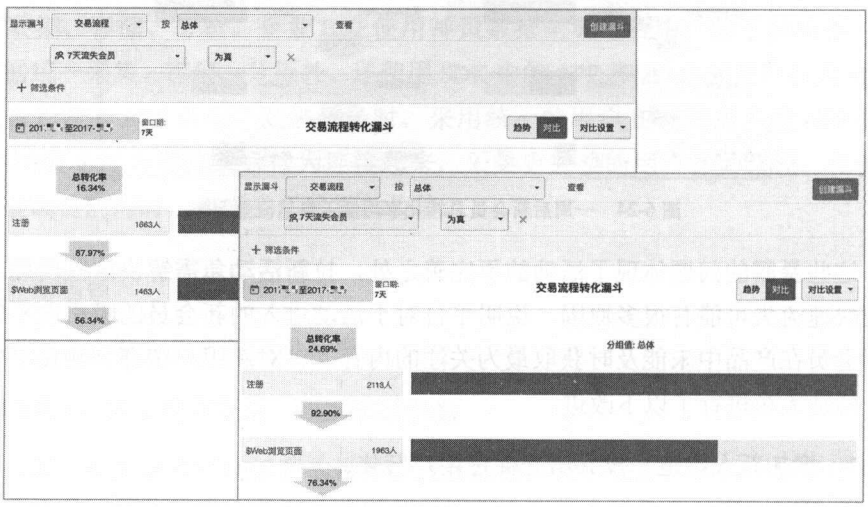


图 6-26　对比精准推送前后会员转化效果

场景 2：可视化全局点击，优化用户体验

着陆页的主要作用在于吸引用户，促进用户继续发生操作。如何优化用户进入产品的着陆页，提升流量转化率，最大化营销效果是企业 B 想要解答的问题。

在电商网站中，首页和商品详情页非常重要，企业 B 结合点击分析功能对着陆页和目标页做了相应优化。

通过点击分析模块选择只查看首日访问用户，了解到产品首页的点击情况如图 6-27 所示。

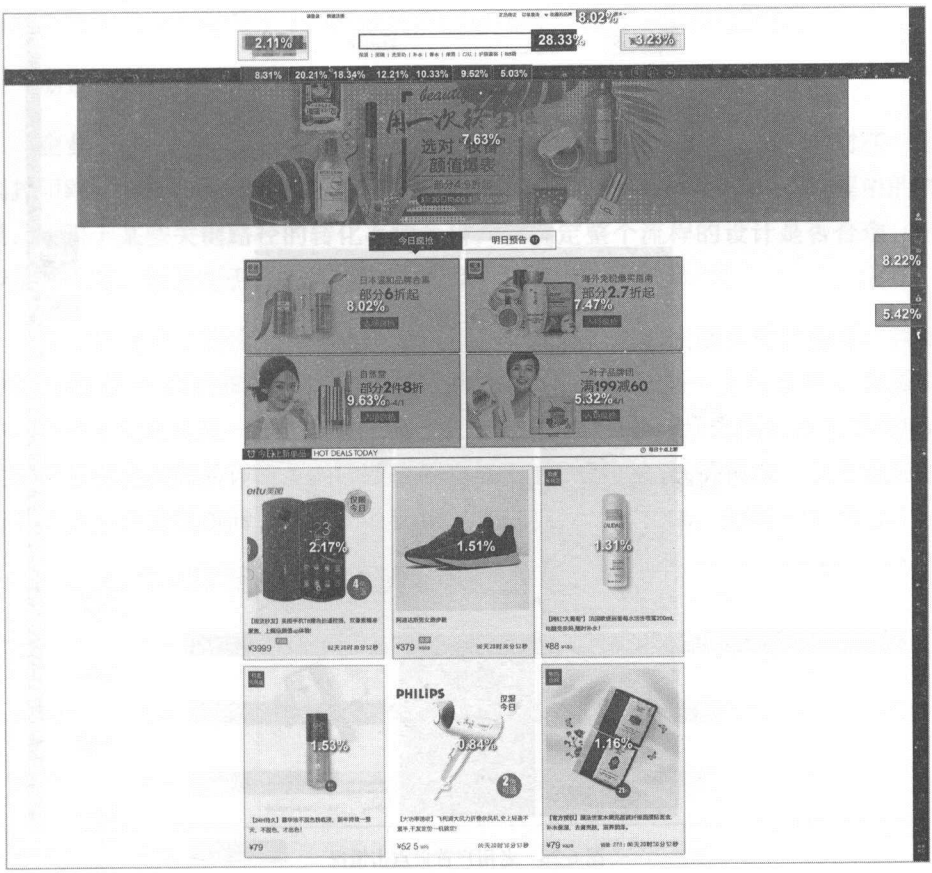


图 6-27　新用户产品首页点击情况

企业 B 发现大部分新用户在首页的点击行为集中在导航栏及搜索框，而其他区域受到了“冷落”。显然，新用户进入网站，很难快速形成对网站的整体认知，或者带着较强的目的性，直接进行搜索。

因此，为了让新用户产生认知和兴趣，企业 B 考虑增加简洁有力的新手引导，并使用更直接的文案，在首页增加优质 UGC 内容的入口，让当下没有购买需求的用户产生好感。同时优化搜索功能，尽可能让有目的的新用户不会失望而归。

同样，图 6-28 显示的老用户首页点击情况显示相比新用户，老用户的点击的元素更多种，行为更密集，多发生在分类专场和个性化推荐版块，而占据最佳位置的 Banner 位点击比例很少。

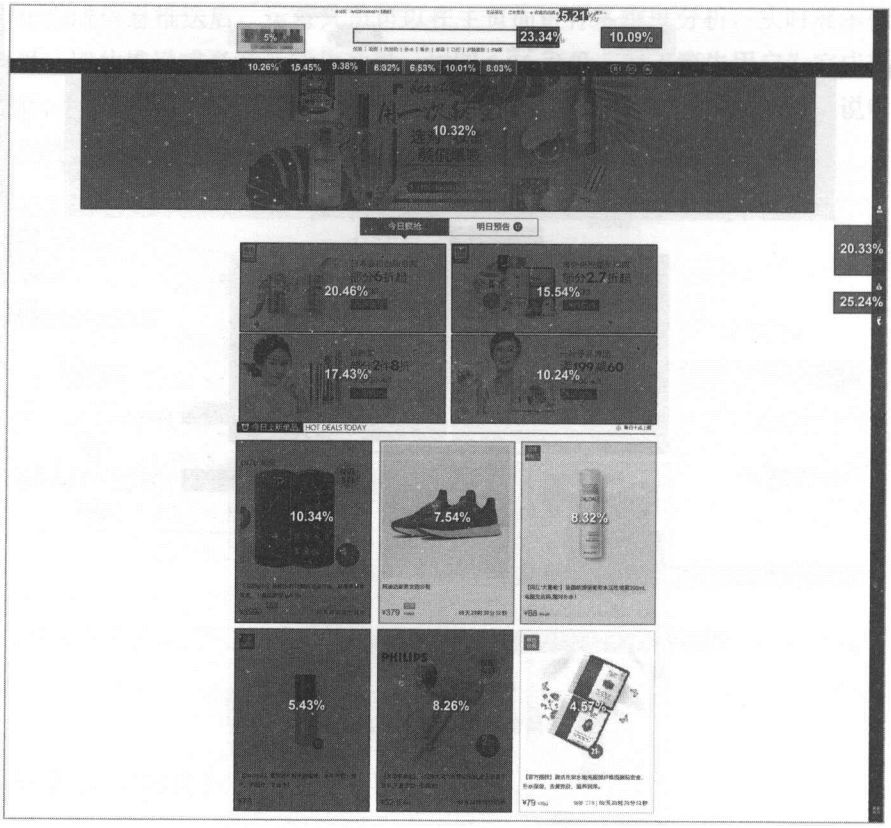


图 6-28　老用户首页点击情况

由于老用户对网站情况较为熟悉，在具体操作方面已经形成固有思维，对于 Banner 位——用来做资源置换或硬性推广的位置，兴趣较弱，因此点击率不高也是意料之中。为了提高整个首页的转化效率，企业 B 优化 Banner 位 UI，引导点击行为的发生。同时在内容方面，提高活动位内容的更新频率，让忠实用户每次访问页面时，都有新鲜的内容可浏览。

从前面的“点击分析”模块的分析，还可以看到老用户在“我的账户”“购物车”“心愿单”等按钮的点击率非常高，这意味着老用户一般直接浏览、挑选自己曾经购买过、或加入心愿单及购物车的商品。针对这一点，企业 B 为用户优化体验流程，在历史订单和心愿单版块中增加高质量的个性化推荐，引导老用户购买新商品。

这一系列优化，将用户从着陆页到目标页之间的转化率提升了 11%，同时，

新用户的次日留存提升了 14%，老用户的使用频次也有一定程度的提高。

场景 3：发现新问题点，搜索贡献增长

企业 B 和大多数电商企业一样，十分注重购买漏斗。漏斗模型是针对在一系列封闭路径中，用于衡量整个转化路径的效率和其中每个步骤的转化过程中的效果。适用于某些关键路径的转化率的分析，以确定整个流程的设计是否合理、各步骤的优劣，以及是否存在优化的空间等。

企业 B 建立了符合购买流程的漏斗，其中有一个是搜索购买转化漏斗，具体步骤为 搜索 → 浏览商品详情页 → 加入购物车 → 提交订单 → 支付订单 。在漏斗中，企业 B 发现从第一步“搜索”到“浏览商品详情页”两步之间的流失率很高。发现搜索模块到商品详情页的转化率明显低于 Banner、精选等模块，且搜索模块带来的商品详情页访问量只占商品详情页总体访问量的 4.5%，如图 6-29 所示。

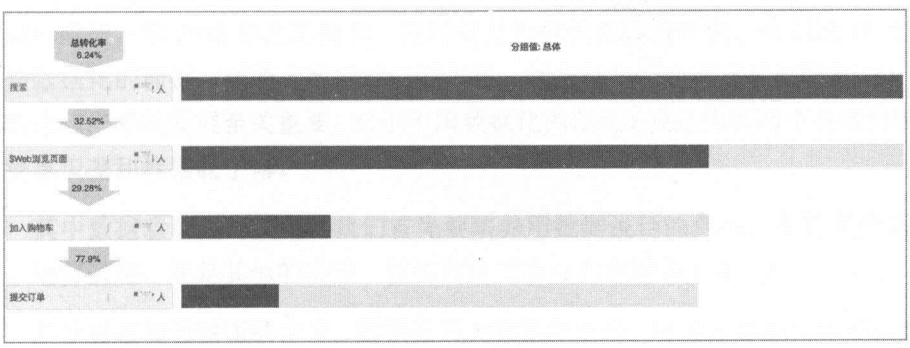


图 6-29　从搜索到购买的转化漏斗

由于之前企业 B 运营人员一直很注重 Banner 位、精选位等位置的优化，未关注过搜索结果的优化，经过此次发现后，产品和运营人员针对搜索模块进行了以下修改。

1. 优化搜索结果内容默认排序规则，增加切换排序规则的功能按钮。

2. 针对没有搜索结果和结果较少的关键词，进行相近内容的推荐商品。

3. 根据用户可能错误的搜索词，给出搜索错误提示。

4. 在搜索中增加高频推荐词模块。

如图 6-30 所示，经过此次针对性优化后，搜索购买转化漏斗第一步转化率得到很明显的提升，总体转化率相比之前提升一倍，同时，商品详情页总体访问量

中，由搜索模块带来的流量由原来的 4.5% 提升至 12%。

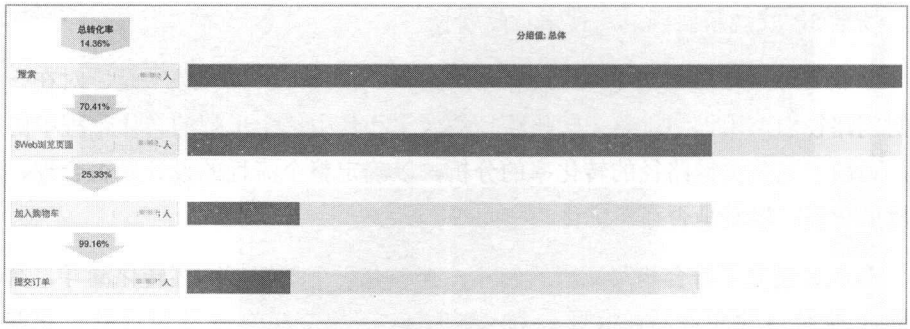


图 6-30　产品改版后漏斗的转化情况

写在最后的话

我个人把2000年后的中国互联网分为两个阶段，其中2000—2015年是IT化建设的阶段，而2015年之后很长一段时间是数据化建设的阶段，可以说IT化建设是数据化的前提。随着大数据时代的兴起，越来越多的企业意识到数据对一款不断迭代产品的发展至关重要。对于中国数据化的现状，我认为有两个关键问题：数据意识差和数据底子薄。

其中数据意识最为关键，我们首先要培养用数据说话的意识，不管是产品改进、运营监控，还是其他的决策，首先问自己有没有数据做支撑。

其次要重视数据源的建设。数据源乃大数据之根基。管理数据源如扎根土壤，根基稳固方能避免“空中楼阁”。这是我在大数据行业工作近十年的最大心得。

最后，驱动决策并非充分发挥大数据的全部价值，让产品智能化更代表行业的发展方向。目前大部分数据分析产品可以满足企业在决策层面的分析需求。在未来，随着大数据在行业应用中的深化，将更加依赖强健的数据仓库和灵活的平台开发能力，通过基础数据叠加算法模型，从而驱动产品智能化。

期待和各位同仁一起，推进中国的数据化建设。

体验神策数据

Broadview®
www.broadview.com.cn
博文视点·IT出版旗舰品牌
技术凝聚实力·专业创新出版

信息随时随刻在产生，它为世界指出两条路：一条路布满着那些故步自封、因循守旧企业的“尸体”；另一条则为拥有数据思维和掌握数据驾驭能力的企业铺就康庄大道。而此时此刻，你正处于交叉路口，手中恰好握着一张指引正确路径的“地图”。

Alistair Croll

哈佛商学院访问执行官，Coradiant 公司联合创始人，《精益数据分析》作者

数据分析是一种修行，“修”是思考的能力，“行”是落实成为方案的方法。文锋在书内的描述正是他这几年创业的发现与精华，值得一阅。

车品觉

红杉中国专家合伙人、全国信标委大数据标准工作组副组长

这是我期待很久的书，本书的结构和内容都经过了反复打磨，无论是从技术严谨性，还是从内容的实用性上看，都堪称互联网商业数据的可贵佳作。

宋　星

互联网数据官创始人、网站分析在中国创始人

书中提到的问题场景，相信也是很多从业者经常遇见的，对于希望提升数据决策能力、了解数据决策真相的从业者，这本书都是很好的读物。

曹　政

曾任百度商业分析部经理，现知名 IT 自媒体博主，互联网游戏出海领域创业者

干货满满，源于实践又高于实践。本书字里行间生动活泼，也体现出作者对大数据领域的理想情怀和脚踏实地的实干家精神，非常值得一读。

吕厚昌

曾任百度高级总监，Pinterest 大数据部负责人

我们相信桑文锋在驾驭数据驱动商业的能力，也相信他身上那股坚定的信念，他愿意花很多年将数据基础能力变成像水和电一样提供给中国企业。我们将自己的资本和信心赌到桑文锋身上。我们也相信这本书，会给希望在商业战场上多一双数据的眼睛的企业家很多帮助。

王　淮

《打造 Facebook》作者，线性资本创始合伙人

上架建议：数据分析

博文视点Broadview

新浪微博 weibo.com
@博文视点Broadview

策划编辑：符隆美
责任编辑：张春雨
封面设计：吴海燕

ISBN 978-7-121-33451-1

9 787121 334511 >

定价：49.00元